Social Network

社会网络

Big Data

大数据 下

企业舆情建模和管理

Enterprise Public Opinion

Management

梁循 杨小平 李志宇 编著

清华大学出版社

社会网络大数据下
企业舆情建模和管理

梁　循　杨小平　李志宇　编著

清华大学出版社
北　京

内容简介

本书综合了国内外的最新资料和作者的研究成果。通过研究社会网络上用户的行为理论，探索适用于社会网络大数据环境下的企业舆情挖掘方法，提出若干个典型的企业舆情发现与合理处置方法。首先提出形式化定义，为后续定量描述用户行为奠定基础，然后进一步应用到企业课题，并研究企业舆情管理优化问题。在基础模型方面，本书研究了基于文本与图像内容的企业舆情模型和基于网络结构的发现模型。在衍生模型方面本书探讨了社会网络用户行为和企业舆情管理优化方法的一些具体课题。

本书的读者可以是对社会计算感兴趣的专业人士，或是对社会化媒体挖掘感兴趣的商业界人士，也可作为计算机应用方向的教材或参考书。

图书在版编目(CIP)数据

社会网络大数据下企业舆情建模和管理/梁循，杨小平，李志宇编著．--北京：清华大学出版社，2016
ISBN 978-7-302-41054-6

Ⅰ．①社…　Ⅱ．①梁…　②杨…　③李…　Ⅲ．①企业管理－公共关系－舆论－研究　Ⅳ．①F270

中国版本图书馆 CIP 数据核字(2015)第 173423 号

责任编辑：刘向威　李　晔
封面设计：
责任校对：李建庄
责任印制：

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编：**100084
社 总 机：010-62770175　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：170mm×230mm　**印　　张：**11.75　**字　　数：**194 千字
版　　次：2016 年 2 月第 1 版　**印　　次：**2016 年 2 月第1 次印刷
印　　数：1～　000
定　　价：　.00 元

产品编号：065993-01

前言

FOREWORD

本书讨论了在社会网络大数据环境下，企业舆情的建模和管理问题。本书与作者先前出版的另外11本书籍《网络金融》、《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》、《电子商务理论与实践》、《网络金融信息挖掘导论》、《网络金融系统设计与实现案例集》、《互联网金融信息智能挖掘基础》、《支持向量机算法及其金融应用》、《金融数据挖掘——基于大数据视角的展望》、《面向社会化媒体大数据的社会计算》、《社会化商务理论与实践》和本书之间的关系见图0-1。

本书的编写得到了中国人民大学科学研究基金项目(10XNI029)的支持。作者的一些学生也参加了本书的编写，这些同学包括张海燕、申华、施晓菁、马跃峰、马超等。

由于作者水平和时间的限制，书中一定存在不少缺点和错误，恳请读者批评指正。

编　者

2015年12月

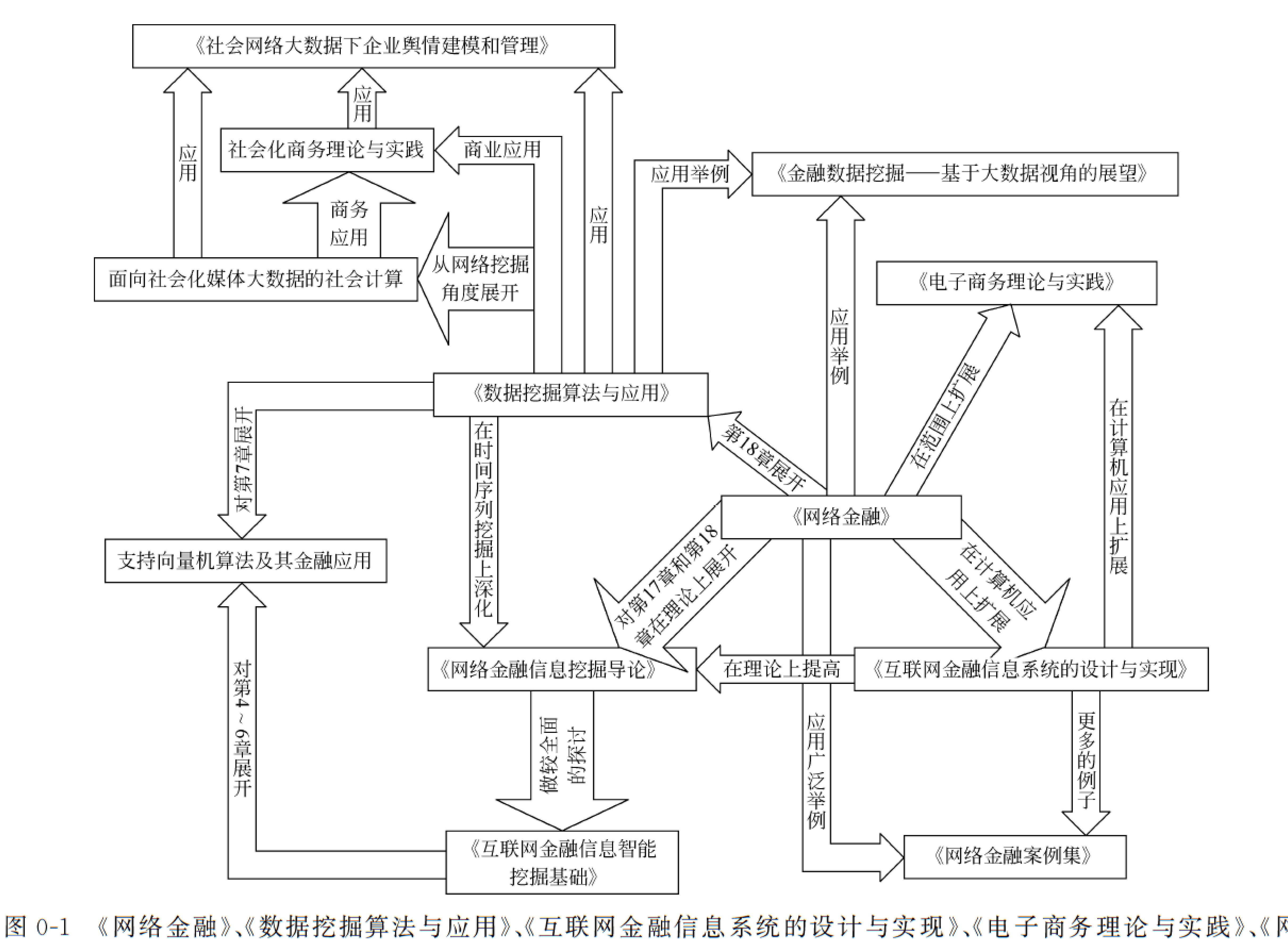


图 0-1 《网络金融》、《数据挖掘算法与应用》、《互联网金融信息系统的设计与实现》、《电子商务理论与实践》、《网络金融信息挖掘导论》、《网络金融系统设计与实现案例集》、《互联网金融信息智能挖掘基础》、《支持向量机算法及其金融应用》、《金融数据挖掘——基于大数据视角的展望》、《面向社会化媒体大数据的社会计算》、《社会化商务理论与实践》和本书之间的关系

目录
CONTENTS

第1章

绪　论

本章学习目标

- 学习社会网络、社会网络大数据、网络舆情与企业网络舆情的概念。
- 理解社会网络计算的研究内容。
- 了解企业社会网络舆情的特点。

1.1　社会网络

近年来，在新的功能、技术和标准的推动下，网络变得更加社会化和互联化，社交性网络平台得到了前所未有的迅猛发展。这些网络平台（包括移动网络平台），如，Facebook、Twitter、新浪微

博、微信、网易新闻等不仅聚集了大量用户，也为企业提供了获取巨大潜在客户资源的渠道。据美国调查公司 Unity Marketing 调查发现，社交网络用户数量已经占据了互联网用户总量的 40%。

社会网络(social network)是指社会个体成员之间因为互动而形成的相对稳定的关系体系。在社会网络中，人们形成了“社交圈”、“兴趣圈”等关系。从狭义上来说，社交圈是我们日常生活中与朋友、同学、同事之间的各种关系网络构成的一个人际圈子。从广义上来说，社交圈可以延伸为我们每个人的生活圈。

社交圈的存在让营销者更为欣慰。以人人网、Facebook 为例，大多数的注册会员都是来自于各个高校的学生，即便日后工作生活变迁，但是这些稳固的同学关系却不会消失。同样，在以白领为主要客户群体的开心网上，同事间的关系虽然相对于同学来说因为变化更大而淡薄一些，但是依然是一种相对稳固的人际关系(Poyry，2013)。互联网社交圈的建立与发展还取决于社交网站用户的交互方式(Backstrom，2006)。社交圈中不同身份的人的影响力也不尽相同(Stutzman，2006)。

社交圈对于企业进行网络营销来说无疑有着积极的促进作用，无论是品牌的创建或者是促销的推进，在稳定网络中的传播广度与深度，都会比大众网络更有优势。

社交圈的分享也可以分为两种：一种是消费者主动的分享；一种是营销者促使的被动分享。被动分享一般都是由网络营销者来促成的，他们往往会针对消费者、潜在客户的商品促销信息、订

单信息的分享、发布行为给予一定的奖励，以这种激励方式达到自己的促销、品牌传播目的。

兴趣圈从理论上看，其范畴应该要比社交圈更大一些。相对于社交网站的社交圈为主的状况，微博、群组、知识分享平台、视频图片网站、团购、LBS服务等，大多都是基于兴趣形成的社会化的关系网络。当然，广义上来说这种因为兴趣而产生的关系网络也属于社交范畴，但是单纯从社会网络的类别、数量上来区分的话，因为兴趣而形成的社会网络应用种类更多、品种更齐全。

常见的社会网络中，除了社交网站之外，还存在大量的其他网站。比如维基百科、百度百科为代表的知识分享平台，以Twitter、新浪微博、腾讯微博为代表的微博平台，以YouTube、优酷为代表的视频分享网站，以Flickr为代表的视频分享网站，以豆瓣、百度文库为代表的文档分享网站，以大众点评为代表的消费评论网站，甚至包括各种团购网站，以及以微信为代表的移动用户端上的平台，其实都属于兴趣圈网络的范畴。

以兴趣圈形成的网络具备一些特殊的特性，那就是社会网络的自优化特性。以最典型的新浪微博为例，对于一个商家或者公众意见领袖来说，普通的网民与之形成的关注同时是不相互关注的关系，从本质上来说并非是社交关系。同样地，这种大多数时候的单向互动也完全不符合社会网络的互动特性，而只是类似于更多的媒体中心进行的单向信息传递而已(Bouras,2004)。

对于商家来说，交互从来也都不是充分的，而这正是社交圈之

所以存在的原因和价值。社交圈可以分为陌生人购物分享网络,例如蘑菇街、美丽说,以及熟人圈社交网络例如微博、人人网等。对于商家,他们更关注的是这些人是否真正地对他们的商家和产品感兴趣,而对于每一个用户的深度交流,只能作为一个远景却未必能实现。而社会网络活动中的自优化功能会帮助商家找寻到最忠实的用户,也就是长期留存下来的稳定网络。

基于社会计算的领域视角,本章通过一种以 Web 2.0 思想为核心的社会计算模式对社会网络进行剖析。因为社会计算及社会网络都是相对较新颖的概念。本小节重点对社会计算和社会网络的相关概念及研究成果做出较为完善的阐述。

1.2 社会网络大数据

在实践中,社会网络的数据量非常大,形成了大数据。大数据指的是所涉及的资料量规模巨大,大数据这个术语最早期的引用可追溯到 apache org 的开源项目 Nutch。当时,大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。随着谷歌 MapReduce 和 Google File System(GFS)的发布,大数据不再仅仅用来描述大量的数据,还涵盖了处理数据的速度。

最早提出大数据时代到来的是全球著名管理咨询公司麦肯锡,麦肯锡称:“数据,已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素。人们对于海量数据的挖掘和运用,预示

着新一波生产率增长和消费者盈余浪潮的到来。”

大数据一词由英文“Big Data”翻译而来，过去常说的“信息爆炸”、“海量数据”等已不足以描述这个新事物。全球著名管理咨询公司麦肯锡的报告《大数据：创新、竞争和生产力的下一个前沿》对大数据做了如下定义：大数据是指大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群。

大数据在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，近年来，随着互联网和信息行业的发展，大数据引起了更多人的关注。大数据在互联网行业指的是这样一种现象：互联网公司在日常运营中生成、累积的用户网络行为数据。这些数据的规模是如此庞大，以至于不能用 GB 或 TB 来衡量，大数据的起始计量单位至少是 PB(1000 个 TB)、EB(100 万个 TB)或 ZB(10 亿个 TB)。

在信息技术不发达的年代，存储设备的价格昂贵，数据的保存所付出的代价是非常大的。随着科技的进步，存储设备便宜了，数据可以在较低的成本下得到妥善的保存，用户自己产生的数据得到了重视。也正是由于数据的价值被重视了，因此越来越多的数据被持续保存，大数据由此产生。网络数据的十几年的保存产生了大数据。

大数据的首要特征是数据量大。截至 2009 年，美国几乎所有部门中每个雇员数量在 1000 人以上的企业所存储的数据平均值至少为 200TB，是美国零售商沃尔玛 1999 年数据仓库的两倍。很

多经济部门中，每个企业平均存储数据超过 1PB。欧洲组织 2010 年存储容量总计接近 11EB，大约为整个美国存储容量的 70％。全球企业 2010 年在硬盘上存储了超过 7EB 的新数据，消费者在 PC 和笔记本电脑等设备上存储了超过 6EB 新数据，而 1EB 数据就相当于美国国会图书馆中存储数据的 4000 多倍。数据容量增长的速度大大超过了硬件技术的发展速度，以至于引发了数据存储和处理的危机。大量的数据会被处理掉，比如医疗卫生提供商会处理掉他们所产生的 90％的数据（包括手术过程中产生的几乎所有实时视频图像）。

此外，大数据不只是大。海量数据引发的危机并不单纯是数据量的爆炸性增长，还牵涉到数据类型的改变，也称为多样化。原来的数据都可以用二维表结构存储在数据库中，如常用到 Excel 软件所处理的数据，称为结构化数据。但是现在，更多互联网多媒体应用的出现，使诸如图片、声音和视频等非结构化数据占到了很大比重。有统计显示，全世界结构化数据增长率大概是 32％，而非结构化数据增长率则为 63％。预计未来用于产生智慧的大数据，往往是这些非结构化数据。

1.3　社会网络计算

社会网络计算，也称社会计算，是对社会网络的智能计算。它作为一个新兴跨学科的研究领域，目前还没有一个公认的定义。

不过，我们可以从社会计算出现的背景去剖析概念，将社会计算概括为“用社会化方法计算社会”，具体包含两层意思，即“为社会计算”和“用社会化方法计算”。

所谓“为社会计算”反映了社会计算研究与服务的对象是社会，包括虚拟网络社会和现实社会，以及从中抽象出来的人工社会。从这个角度来说，通过信息技术方法对社会数字轨迹进行分析，了解社会已经发生，监控正在发生和预测将要发生的事情，准确地把握社会的动态特征和运行规律，预测政策实施的可行性，为虚拟网络社会的科学管理和政府决策提供有效参考。

所谓“社会化方法”就是以“草根”客户为中心，并依靠“草根”客户的方法，是一种协同和群体智能的方法，一种从个体到整体，从微观到宏观的思维模式。许多事件都是由无数网民的“你一言我一语”和微不足道的微观行为最终发展成为一个重大的社会事件或浩大工程。如维基百科就是由无数网民微不足道的努力而完成的巨大的百科全书，这靠少数专家是无法完成的。从这个角度来讲，社会计算是一种群体智能的计算模式（Liang，2012；梁循等，2014）。可以看出，社会计算本质就是对社会网络进行智能分析的过程。

社会计算的研究对象是社会，包括现实的物理社会和虚拟的网络社会（Moreno，2004；Cebi，2013）。前者主要指传统意义上的社会，如某国家或地区；后者主要指基于 Web 的虚拟网络社区。从广义来讲，整个 Internet 就是一个虚拟网络，但从狭义来讲，虚

拟网络主要指基于 Web 2.0 的,强调以客户为中心的虚拟社区,如 Facebook、Twitter 等虚拟网络。但需要指出的是,尽管社会形态可以分为现实物理社会和虚拟网络社会,但两者又是紧密相关的。虚拟网络社会是对现实物理社会的反映,研究虚拟网络社会的最终目的还是为现实物理社会的管理服务。

从微观客户层面来讲,社会计算主要关注的是如何促进客户与客户的交互,以及通过客户交互表现出来的客户社会影响分析。

(1) 客户交互研究。无论是 Web 2.0 还是 Facebook、Twitter 等虚拟社会网络系统,其最大的特点就是强调客户与客户间的交互,实现的是人与人的互联。如何促进人与人的交互是社会计算研究的另一重要内容。一般认为,随着 Web 2.0 理念的深入,交互的重点已经从传统的人-机交互(human computer interface,HCI)转化为人-人交互(human human interface,HHI)。传统人-机交互强调的是通过设计人员对系统形式和功能的控制来优化软件应用及界面以增加系统的友好性,而人-人交互更注重如何实现人与人互联,信息交换与知识共享(毛基业,2011)。对不同的应用领域,人-人交互的模式不同,如在微博中,交互方式包括跟帖、回复、粉丝等;在微信等人际关系网中,人-人交互一般显性表现为加某某为好友。

(2) 客户影响分析。通过客户间的交互(回复、跟帖、加为好友等),客户与客户间形成一定的影响关系,并会逐步形成社会网络中的影响力(Ghosh 和 Lerman,2010)。客户影响力分析主要研

究如何基于客户的交互活动水平(activity level)来研究客户与客户是如何影响的,以及客户在社会网络中的影响力大小。具体来讲,影响分析包括影响关系分析和影响力分析。影响关系分析是划分客户间影响关系的研究,影响力分析用于度量客户在社会网络中的重要程度。有较多学者基于博客、论坛行为(提交、评论等),采用 Web 挖掘的方法和语义分析的方法研究博客、论坛客户的影响力。从模型来讲,度量客户影响力通常表现为寻找最重要的节点,目前主要有两类:第一类是基于最小路径(geodesic path)的方法,如距离中心(closeness centrality)、图中心(graph centrality),中介中心度(betweenness centrality)等;第二类是基于拓扑结构的方法,包括基于马尔柯夫的方法(如 PageRank 算法、HITS 模型)、度中心(degree centrality)的方法、基于路径的方法(如 α-centrality、SenderRank 等)。

1.4 舆情、网络舆情与企业网络舆情

如果要分析网络舆情的含义,首先要理清楚舆情的概念,因为网络只是舆情传播在现代社会的载体,只是舆情的传播具备了新的特性而已。而对于舆情具体的概念,不同的学者也有不同的看法,目前较为流行的有关舆情的定义主要有以下几种:

狭义舆情的定义,王来华(2003)在《舆情研究概论》中将舆情定义为“在一定的社会空间内,围绕中介性社会事项的发生、发展

和变化,作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度。"虽然作者将民众的意愿限定在民众的社会政治态度方面,但是,它所包含的舆情对于国家管理者的利益关系变得更加突出。

偏重于民意的定义,张克生(2004)在其专著中将舆情定义为"国家决策主体在决策活动中必然涉及的、关乎民众利益的民众生活(民情)、社会生产(民力)和民众中蕴含的知识和智力(民智)等社会客观情况,以及民众在认知、情感和意志基础上,对社会客观情况以及国家决策产生的主观政治态度。"

较为全面的定义则是刘毅(2007a)在《网络舆情研究概论》中表述的:"舆情是由个人以及各种社会群体构成的公众,在一定的历史阶段和社会空间内,对自己关心或与自身利益紧密相关的各种公共事务所持有的多种情绪、意愿、态度和意见交错的总和。"因为民众的舆情所指并不一定都指向国家管理者,如果具体到企业舆情更是如此,例如马航事件,包含民众的一种关怀与"泛亲情",也包含对马航公司的谴责等等,可以说是同时多种意见与情绪的集合。

由此,我们认为网络舆情是社会舆情的一种具体表现形式,是公众在 Internet 上公开表达的对某种现象、问题或具体事物的具有一定影响力和倾向性的共同意见的集合。而企业网络舆情则将网络舆情进一步缩小,限于企业这个主体相关的舆情内容。因此,我们所使用的企业网络舆情定义为:"由个人以及各种社会群体

构成的公众，在一定的历史阶段和社会网络空间内，对自己关心或与自身利益紧密相关的特定企业的产品、经营和管理，通过网络表达的多种情绪、意愿、态度和意见的集合。”企业网络舆情借助于网络的传播，是网络舆情在互联网空间的进一步降维映射。

近年来，舆情监控成为国家管理互联网的一种必要技术（Liang 等，2012）。利用计算机智能技术，可以将各种人类情感转化成实实在在的数值型数据。情感倾向性分析比较系统的研究工作，开始于基于监督学习方法对电影评论文本进行情感倾向性分类和基于无监督学习对文本情感倾向性分类的研究（王超等，2009；桂斌和杨小平，2015）。情感倾向性分析也称为情感分类、情感分析、文本意见挖掘、观点挖掘等，涉及自然语言处理、信息检索、数据挖掘等研究领域。一般分为文档级观点挖掘和语句级观点挖掘，其情感倾向包括简单的赞同、反对、中立三种态度，也包括对某一对象所持态度的强度，甚至舆论对该对象的具体看法和态度等。情感倾向性分析目前已经获得了很大程度的发展，特别是在线评论的情感倾向性分析（叶强，2007）获得了很大的发展。目前基于在线评论文本的情感倾向性分析的准确率最高能达到 90％以上（杨源等，2012；施晓菁和梁循，2015）。为了找出评论的情感倾向性，我们需要借助的智能手段包括自然语言处理、机器学习、文本挖掘等，从而实现利用计算机自动地识别出互联网文本的感情取向（梁循，2006；Liang 等，2008；Liang，2010；He 和 Zhou，2011；Neviarouskaya 等，2011；Liang 和 Ni，2011；胡百精，2013）。

传统上,公司会为了了解这些信息进行用户问卷调查,这需要花费很多人力对用户满意度进行调查和对问卷进行分析。这种调查的有效性通常是很有限的,原因是调查样本大小的限制和制造有效的调查问卷表的困难。如果能够通过在线文本(如 Web 网页、聊天室和新闻文本)的内容分析,自动探测和分析对感兴趣话题的“喜爱度”,人们就可以很容易地识别这些在线文本中的自然的评价。

1.5 企业社会网络舆情的特点

随着社会网络影响力的增强,越来越多的公司开始关注企业自身在社会网络交互口碑舆情传播中的重要作用(Ye,2012)。目前社会网络大数据环境下,企业社会网络舆情的主要传播渠道包括网络新闻媒体、网络博客、社会网络平台(微博、微信、QQ 等)以及在线商品销售平台,企业社会网络舆情的特性也与其传播方式息息相关,主要包括以下几个方面(刘毅,2007b;吉祥,2010;康伟,2012)。

(1) 虚拟性:网络空间的虚拟性打破了传统的物理空间的界限,是一个无形无界的信息空间。信息的提供者、传播者和阅读者之间的角色没有明显的界限,逐渐形成一种“真实而又虚拟的沟通系统”。

(2) 实时性:网络舆情的实时性主要体现在信息传输的速度

上，由于现代技术的应用以及手机、pad 等移动用户端信息推送功能的增强，在社会网络中进行信息的传播与散布成本大大降低，同时传播的速度也极快地增加。

（3）交互性：交互性是指网络参与主体利用互联网可以通过实时交互操作的方式发表、传播和反馈各种媒体信息。网民利用网络普遍表现出强烈的参与意识，可以不受时间和空间的限制进行交流。

（4）易感性：在网络社会中，由于生活圈子狭小，一旦人在意见上陷入孤独，往往也意味着他在其他方面也陷入孤独，从而产生"从众心理"。网络空间使人们的交往范围大大扩大，人们总能在广阔的虚拟空间中找到拥有同样价值观、兴趣和关注点的人群，这就是网络舆情的易感性。

（5）丰富性：网上舆情的主题极为宽泛，话题的确定往往是自发、随意的。从舆情的话题来看，涉及政治、经济、文化、军事、外交以及社会生活的各个方面；从舆情来源上看，网民可以在不受任何干扰的情况下预先写好言论，随时在网上发布，发表后的言论可以被任意评论和转发。

（6）开放性：互联网采用开放的网络结构，使得企业社会网络舆情的传播平台也具有开放性的结构，这便决定了企业社会网络舆情传播方式的开放性。各种网络交流平台为受众提供了信息共享和互动的渠道，它打破了人与人之间交流的时间和空间的限制，实现了人们能够随时随地进行交流，扩大了信息共享的范围。

（7）突发性：网络舆论的形成非常迅速，这是传统媒体无法比拟的，一个热点事件的存在加上一种情绪化的意见，就可以成为点燃一片舆论的导火索。网络参与主体之间很少进行有效的沟通，某一事件发生时，网民可以立即在网络中发表意见和观点，这些来自不同地方的个体意见可以迅速地在网络平台上汇聚起来形成公共意见。各种渠道的意见又可以迅速地进行互动，从而迅速形成强大意见声势。

1.6 企业社会网络舆情与国家层面舆情的联系与区别

企业级的网络舆情与国家级的网络舆情存在着密切的联系，但同时也具有多种区别。

1. 企业网络舆情与国家网络舆情的联系

（1）在目标上：都是为了建立正面形象，降低负面舆情对自己的影响。

（2）在处理技术上：都是采集网络信息，使用计算机智能数据挖掘技术进行处理。

2. 企业网络舆情与国家网络舆情的区别

（1）在处理技术上：企业舆情管理是针对社会网络上和企业相关的舆情进行的，很大程度上取决于企业自身的利益，而较少关心其他不相关行业企业的信息。从计算机文本智能处理角度看，

其专业词汇有一个范围，或者说可以大体地构造出一个《企业词典》。而国家级的舆情管理是对所有国家层面上的，很难构造出一个词典。

（2）在网络营销上：在经济社会中，企业对社会的经济发展起着至关重要的作用。在企业管理中，各项内容都是在围绕着盈利目标进行的，其社会责任、社会贡献是为树立企业的正面形象，从而间接地支持了其盈利活动。所以，利用好社会网络，建立优质品牌效应，通过和客户在微博等网络上的一对一服务，包括营销服务，进行高质量的客户关系管理，可以直接促进其销售。而国家级的舆情管理则没有产品营销的任务。

（3）在重大舆情事件的发生频率、管理难度和管理成本上：对于企业舆情管理，因为其本身的范围小，所以会对企业产生重大影响的舆情事件发生频率较低，相对较容易管理。而国家级的舆情管理因为其本身边界很大，往往重大舆情事件发生的频率较高，舆情管理需要投入更大的成本。

（4）在语料的积累和交互历史的积累上：由于在社会网络中企业与客户的交互往往较多，如产品出现问题往往有很多网民可以很容易且精确地找到对应企业的官方微博进行交互，所以有较多的语料积累，而政府级的舆情管理往往缺少相应语料，因为当舆情事件发生时，大家往往不知道该去哪里留言。

（5）在舆情管理的时效性上：由于企业的舆情通常集中于短时间内大规模的传播与爆发，因此留给企业进行舆情管理与控制

的时间极短，并且这种影响会立刻显现在企业的销售业绩以及股票价格上（通常在24小时内），因此企业对于舆情管理的灵活性和时效性要求较高。而国家的舆情管理通常侧重在一段时间内对舆情的正确引导，并且所触及的范围更大，影响力更加平均，因此相对来说时效性的要求也就不如企业的舆情管理那么高了。

（6）在应对措施上：从舆情应急处理上看，国家舆情应急管理可以介入国家行政手段干预；而企业舆情应急只能通过网络公关等进行舆情引导。企业舆情管理的应对策略主要包括积极性和消极性策略两种。积极性策略包括及时发布更正，辟谣或纠错信息、材料或证据，企业领导人出面表态或说明，通过第三方公证或表态，积极配合政府调查，向受害者、消费者等群体道歉或补偿，召回部分有缺陷产品，停产整顿，采取赔偿、管护等补救措施，全面改进产品或服务，设立体现企业社会责任的基金等。消极性策略如利用金钱或关系谋求删帖、封帖、换帖等，以使得不利于企业的舆情无法发布或传播，管控、威胁、收买或打击相关受害者或当事人，逼迫或要求相关当事人不再发表不利于企业的言论，回避事件焦点问题或议题而宣传其他信息，否认相关产品或服务有问题，有意将问题或矛盾重点转移到不重要的问题或要素等。而国家级的舆情管理的应对策略则是对舆情的信息收集、分析及其调控政策等。对国家利益受到重大影响的信息，直接让相关网址删除。此外，国家级舆情监控还有预防措施，可以对主要社会网络提出要求，对涉及国家安全的帖子，设一些保护措施，有恶意的用户根本上传不上去。

（7）在靶向目标方面：企业舆情管理有较强的靶向性，主要针对的是与企业相关的人群，包括客户、主要竞争者、上下游伙伴和潜在目标客户等。其舆情管理的目标人群与企业生产、运作、营销、售后等有着强烈的相关性。而在国家级舆情管理的目标人群在深度和广度上都与企业级的不同。一方面，国家级的舆情管理针对的是整个国家的、全方位的一个舆情管理，与企业相比广度更宽，靶向性较企业级弱；另一方面，国家级的舆情分析需要对于社会事件（如公共安全事件等）做出一个全方位的了解、定位和掌控的反应过程，因而其又具有较深的深度性特点。

（8）在舆情的处理方式上，企业舆情管理偏向于运用和发挥，国家级舆情管理偏向于监督和防范。具体来说，企业舆情一般是在充分研究社会网络的基础上，对品牌口碑的研究、竞争对手动态、行业状况分析、热点事件判断等内容进行信息的收集提取、汇总分析及进一步的扩散传播，进而达到其商业盈利的目的。而国家级的舆情管理则注重利用社会网络及时发现社会中存在的不稳定因素，在第一时间进行监管和处理，把危害到社会公共安全的可能降到最低。

1.7　企业社会网络舆情的研究意义

社会网络企业舆情的研究从企业和用户两方面都具有很深远的意义。

从企业的角度来讲，基于社会网络的企业舆情可以帮助企业实现销售业绩的提升，进而提升企业的利润。

企业社会网络舆情也促进商务模式的变革。社会网络可以优化电子商务销售活动，通过对企业社会网络舆情的研究，通过合适的科学方法，找到舆情反映良好的作为正面反馈加以推进，找到舆情反映不好的加以改正。同时，社会网络能够推进交易量，提升转换率并增加在线零售商的平均订单额。例如，通过将他们的触及范围从电子商务网站延伸到社会网络平台，零售商们可以在消费者经常出现的地方开店。诸如 Bazaarvoice、Power Reviews 等社会网络软件供应商已经积累了很多引人注目的数据来佐证将社会网络工具和内容加入或链接到他们自己的电子商务网站的销售价值。举例来说，通过将用户评论加入到它的网站，英国的零售商 Argos 的转换率增加了 10%；通过在社会网络上植入视频广告并将它们链接到自己的电子商务网站上，美国搅拌机制造商 Blendtec 的销售额增加了 500%，同时 Juicy Couture 的在线 Club Couture 社区的转换率提高了 162%。

社会网络是一种商务模式的变革，越来越多的客户被纳入了企业舆情的视野之内，从而提供了更多的机会。社会网络给企业提供了进行商务模式革新的机会——通过组织和提取社会网络内容来创造新的收入来源。例如，在 B2B 部门，一个称为 Westlaw 的公司销售一种包含组织用户内容的有偿标记服务的产品——Peer Monitor，倘若这样，来自合法公司的自定义财务和运营信息，

可以匿名化并将其计入基于地理位置的竞争力表现报告中。在B2C中,耐克和苹果同样地实现了Nike+的服务,在线记录跑步者的数据并提供协作工具和竞争力表现报告。

从用户的角度来讲,社会网络提供信任、实效和趣味,这些都需要被科学高效地确认和纠偏,所以舆情研究意义颇深。

在我国社会文化特征条件下,通过推进社会化互动和电子商务相关站点的用户贡献,比如允许用户评级和评论,对用户而言,网站会变得更具吸引力。这是因为社会网络内容增加了销售和市场信息的资源可信度,使它们对用户而言更可信,更具说服力,也更值得相信。反之,如果不信任了,那么企业就必须做出相应的举动。

社会网络具有时效性,只有科学的方法才可以高效正确地做出最正确的决策。社会网络的工具,比如允许人们在一起分享在线购物行为的社会化购物信息,对于线上消费者是有帮助的社会网络网站,它允许消费者通过同步浏览、社会化书签和团购等工具更精明地购物。通过在消费者常去的地方加入这些工具,品牌商、企业和零售商们能够增强消费者在线购物体验。

社会网络具有趣味性,然而这种趣味应该怎样正确地体现却需要更科学的研究提供决策支持。社会网络的功用除了为用户提供产品发现、选择和参照外,也许还有情感价值——将用户的在线之旅提升为更自然的参与满足感和更有价值的社会化体验。从历史角度看,商业总会贴着社会化的自然属性——共同交易、共同购

物。与此相反,早期电子商务就是一段孤独无趣的经历——人们只能跟软件交互,社会网络再现了社会化的贸易。

1.8 本章小结

社会网络是社会个体成员之间互动而形成的相对稳定的关系体系。狭义上,社交圈是日常生活中朋友间、同学间及同事间的各种关系网络构成的一个人际圈子;广义上,社交圈可以延伸到每个人的生活圈。社会网络的数据量巨大,形成了大数据。大数据的规模一般是PB(1000个TB)级以上。社会网络计算,也称社会计算,是社会网络大数据的一种智能计算。社会网络舆情大数据,也是社会计算的研究范畴。企业级的网络舆情与国家级的网络舆情,在处理技术上、管理成本及舆情处理方式等方面,既存在着密切的联系,也具有多种区别。研究社会网络企业舆情对企业管理有很重要的现实意义。

思考题

1. 简述社会网络与社会网络计算的概念。

2. 总结归纳社会网络大数据的内涵与外延。

3. 简述企业级的网络舆情与国家级的网络舆情的联系与区别。

第 2 章

互联网舆情分析的主要技术

本章学习目标

- 掌握舆情信息抽取方法、摘要提取技术。
- 掌握关键词提取技术及文本倾向性分析技术。
- 了解关联分析技术、主题检测和追踪、舆情热点发现和监测方法。

2.1 引言

互联网目前已经形成一个虚拟社会，这个虚拟社会具有虚拟性、隐蔽性、发散性、渗透性和随意性等特点。如今，互联网已经成为舆情产生和传播的主要场所，并且在社会生活中扮演着越来越

重要的角色。网络舆情的爆发将以“内容威胁”的形式逐渐对社会公共安全形成威胁。加强互联网的管理和监控，展开互联网信息收集整理与挖掘已经成为目前各级政府部门亟须解决的现实问题。

舆情分析的主要内容有很多，典型的包括热点信息的发现和文本倾向性情绪挖掘，例如，热点信息异常发现、基于情绪分析的文本态度挖掘；网络话题自动发现算法，用于解决主题检测技术在现实应用中面临的问题；面向话题的多文档关键词和摘要提取算法，用于自动提取网络话题的关键词和摘要，帮助用户快速了解网络话题内容，分析网络话题的传播趋势、动态演化规律，并用动态传播图的形式展现话题传播的线索，以波谱图的方式展现一定周期内的话题动态演化情况。

国内外越来越重视对于网上舆情的研究，其中涉及的技术种类很多，网络话题发现和分析技术在学术界与工业界已受到较长时间的关注。其中，网络话题的自动发现主要采用主题检测与追踪技术；网络话题的深入分析功能，则包括网络话题的关键词和摘要提取、态度倾向性分析、传播分析、动态演化分析、趋势分析和关联分析等。另外，在舆情分析过程中涉及文本检索的相关技术，国内外的文本检索发展一直较快，尤其是基于关键词的检索。下面将阐述有关技术的发展现状与趋势。

2.2　舆情信息抽取

来自互联网的数据往往具有噪声并且随机性很大的特点，所以从互联网抓取的数据需要进行预处理，包括从 HTML 中提取需要的文本数据、还原或者缩写短语以及一些语法解析工作等等。除此之外，分类算法需要相应的数据库作支持，这些都需要提前准备。领域词库和语法库是最为重要的两个数据库，前者是进行文本解析的基础，存放了大量的专业领域词汇以及舆情倾向性信息，后者存放着分类规则和函数，通常通过训练过程得到。

不同的分类算法在性能和精准性方面各有优劣，我们倾向于首先通过多种分类算法对文本按照舆情分类，之后再利用一种机制将所得到的不同的分类结果综合起来作为最终的分类结果。

目前，以下几种文本分类算法在学术界比较有效，它们分别建立在不同理论和概念基础上。

(1) 基于词频的分类算法。这种分类方法简单直观，并且不需要训练。领域词汇数据库中的每一个词汇都被赋予一个数值，然后根据这个数值对文本中所有出现的领域词汇进行统计，得到一个最终的值，这个值代表了该文本的舆情值。根据这个数值将该文本归为正向、负向或者中性。

(2) 基于向量距离的分类算法。将每一个文本用一个文本向量表示，维数为 D，其中 D 为领域词汇数据库中词汇的个数。对于

文本向量中的每一个元素，其值对应着该词汇在该文本中出现的次数，因为每一个文本中出现的词汇仅仅占词汇数据库中的一小部分，大多数文本向量是稀疏的。这种分类算法需要训练。首先按照以上的规则为每一个训练样本集合中的文本计算出一个向量，然后得到一个向量集合 G，其中 $G=\{g_1,g_2,\cdots,g_t\}$，$g_i(i=1,2,\cdots,t)$代表每一个训练样本的文本向量，其中 t 为训练样本的个数。计算完训练样本的文本向量之后，采用同样的手段计算测试样本的文本向量。

（3）基于强度的分类算法。文本舆情对应着该文本所体现的情感方向，即是正向、负向或者是中性，而文本强度则体现了文本的影响力。强度高的文本具有更大的事件影响力，相反，强度低的文本对于事件而言影响力比较弱。如果赋予每一个文本一个值，那么舆情代表该值的符号，而强度对应其绝对值的大小。

这里采用的分类算法就是通过统计不同词汇的强度来得到最终文本的强度。通过

$$F(w)=\frac{\frac{1}{|C|}\sum_{i\neq k}(u_i-u_k)^2}{\sum_i \frac{1}{n_i}\sum_j (m_{ij}-u_i)^2},\quad \forall w$$

来计算每一个词汇的强度，其中 i 用来索引舆情类别，这里 i 的取值可以是 1～3；j 用来索引文本；w 代表词汇数据库中的一个关键词；C 代表舆情类别的个数，这里 C 的取值为 3。以上的式子要在训练样本上进行计算，训练样本的舆情类别已经为其标记。对

于一个特定的词汇 w 而言，u_i 代表它在训练样本类别 i 的文本中出现的平均次数，m_{ij} 代表它在类别为 i 的文本 j 中出现的次数，n_i 代表该词在类别 i 中总共出现的次数。因此可以得到，对于每一个词汇 w 而言，计算的是其在类别间的出现次数变化和类别内的出现次数变化的比值，这在一定程度上反映了该词在舆情类别上所体现出的强度。比值越大，说明该词的强度越高。于是，基于这个机制借助训练样本集合对每一个词汇都计算出一个强度值，然后将得到的信息用于计算测试样本的文本强度上。正像前面所说的，舆情类别反映了强度的符号，将正面的词汇计算成正值，负面的词汇计算成负值，中性词汇值为 0，最后根据每一个文本计算出来的值将其按照舆情进行分类。

2.3　关键词提取

对于关键词提取，tf-idf（term frequency-inverse document frequency）是一种经常被提及的概念。tf（term frequency）代表“关键词的频率”或者“单文本词汇频率”，即某一文档中某一词条出现的频数，tf 越高意味着更强的区分文档内容属性的能力，其权值越高。idf（inverse document frequency）代表“逆文本频率指数”，简称倒序索引排序，即文档集中包含某一词条的文档数，idf 越高说明它区分文档类别属性的能力越低，其权值越小。

所谓关键词，即在单一文档中出现较多同时在其他文档中出

现较少的词，也就是 tf×idf 值比较大的词。所谓关键词提取，其实就是选取 tf×idf 值比较大的词作为关键词的过程。这里的“比较大”既可以是相对的概念，即选取权重排名前一定比例的词作为关键词（如前 10％），也可以是一个绝对的概念，即规定关键词的数量，如权重排名前 100 个词作为关键词。为文档提取关键词，大大减少了信息的冗余，为后续的智能决策提供了便利，具有很重要的实际意义。

2.4 摘要提取

关于摘要提取的研究，数字图书馆领域的著名国际会议——ACM/IEEE 联合数字图书馆大会（JCDL）、国际计算语言学大会（ACL）、国际信息检索大会（SIGIR）上均发表有关文档自动摘要的最新成果。国际上文档自动摘要方面比较著名的几个系统包括 ISI 的 NeATS 系统，哥伦比亚大学的 NewsBlaster 系统，密歇根大学的 NewsInEssence 系统等。国内的哈工大信息检索实验室，清华大学智能技术与系统国家重点实验室等都对此问题进行了一些研究，但没有看到成熟的系统。国外的文档理解大会（DUC）专注于文档摘要的评测，包括单文档摘要、多文档摘要、主题相关的多文档摘要等任务。国内的基础资源与评测也进行了单文档摘要的评测任务，但测试集规模比较小。

文档摘要技术的研究在图书馆领域和自然语言处理领域一直

都很活跃，最早的应用需求来自图书馆。图书馆需要为大量文献书籍生成摘要，而人工摘要的方式效率很低，因此亟须自动摘要的方法取代人工，高效地完成文献摘要任务。随着信息检索技术的不断发展进步，文档自动摘要在信息检索系统中越来越重要，逐渐成为信息检索领域的研究热点之一。

目前，业界出现了一些文本挖掘产品，能够提供单文档摘要功能，例如拓思尔、海量科技公司等的产品。百度搜索和纳讯新闻搜索都能为检索到的文档提供简单的单文档摘要。文档摘要为主题检测与追踪提供服务，通过摘要信息方便对主题内容进行了解，可提高检索效率。

已有的摘要技术大都基于与主题无关的摘要方法，针对候选的文本进行后置处理来判断主题。这种判定的方法没有充分利用到文本内部的关联关系，因此摘要的准确性不高。国内针对中文文本的摘要方法大致分为两类，分别是基于相邻段落语义相似性的方法和基于篇章结构图的方法。前一种方法认为一篇由多个段落组成的文本，其相邻段落在内容上是相近的，形成多个语义内聚的节。通过分析相邻段落间的语义关系即可自动地实现文本主题的划分，两个相邻段落间的语义关系通过它们所共有的词条数来衡量，在处理篇章结构比较规范的文本时效果很好。然而，当文本写作风格自由，且主题分布灵活多样时，采用此方法的效果会大打折扣。主要原因在于它仅仅计算了相邻段落间的语义相似性，而忽视了对那些可能会跨段落分布的主题的处理。此外，采用该方

法必须人工来主观设定段落间的语义相似度阈值,而由于阈值的设定往往和文本的题材、体裁等因素相关,因此通过人工来预先设定阈值往往并不合适。后一种方法不同之处在于采用了基于篇章结构图的策略来计算文本中各段落之间的语义距离,以段落为节点,段落间的语义距离为边构造文本的篇章结构图,进而根据人工设定的某语义距离阈值来分析文本中可能包含的潜在主题。

上述方法的共性在于需要采用以句子作为基本单元的抽取策略,摘要长度无法确定,需要事先直接或间接的给定,实际情况是,不同文本的主题分布灵活多样具有截然不同的信息量,为了保证摘要产生的准确全面而且不失简洁,需要能够通过文本深度挖掘的办法寻找潜在的主题相关的内容,对文本句子的关系进行处理,进而产生更准确的摘要。在关键词自动提取部分,需要考虑关键词词频的历史波动,对于热点事件的检测十分有利,能够更加客观地反映当前关键词汇的异常状态。

2.5 文本倾向性分析

文本倾向性分析也称情绪判定。对于互联网文本而言,决定其性质的因素有很多,包括外在因素和内在因素。前者比如文本的数量,即特定时间内互联网上出现的关于某项话题的文本的个数。后者主要描述单个文本的性质,对于单个文本而言,它的性质可以取决于其内容和强度,内容为该文本的主题、时间、文体等,强

度主要指该文本的影响因子,即该文本的出现可以多大程度上对相关领域的人和事物产生影响。

近几年以来,文本舆情分析的研究逐渐成为国内外研究者所关注的一个热点。通俗地说,文本舆情描述的是文本所传递的情感。对文本舆情进行分析,实际上就是试图根据文本的内容提炼出作者的情感方向。但是我们希望这项工作可以由计算机实现。因此文本情感分析是指通过计算机技术自动分析文本信息所包含的情感因素,例如喜欢或讨厌、正面或负面、快乐或悲伤、愤怒和恐惧等。由此可见,文本舆情一定程度上代表了作者的感情取向,决定了文本内容的褒贬含义。为了实现对文本信息的准确提取,我们不仅需要掌握该文本的影响强度,同时还需要对文本的感情取向有一个正确的把握;如果我们需要对每一个文本赋予一个值,那么影响强度可以看成是其绝对值的大小,而舆情可以看成是其正负号。

人可以很容易地对一个文本的感情取向进行判断,但是我们现在希望可以通过计算机手段自动地提取文档中的情感因素,从而实现批量且实时的处理。为了实现这个目标,需要借助的技术手段有自然语言处理、机器学习、文本挖掘、模式识别等等。借助它们,可以利用计算机自动地识别出这些互联网文本的感情取向,即舆情,然后实现对这些文本基于舆情的分类。

对文本中所包含的情感相关内容进行分析和自动计算,在相关研究中通常称为文本情感分析技术,或者叫做文本态度倾向性

研究。文本情感分析有着众多的潜在应用领域,并为自然语言处理提供了新的研究思路和研究角度。

文本舆情分析可以进一步划分为以下几步工作:词汇语义褒贬分析、文本整体感情色彩分析以及文本观点提取和总结。词汇语义褒贬分析是文本情感分析研究的基础,是基于词汇粒度上的舆情分析,包括分析其语义属性和其强度因子,其中主要借助统计方法和语义方法。

文本整体感情色彩分析是在第一步的基础上进一步通过语义方法或者机器学习的方法确定整个文本的舆情。最后,文本观点提取和总结主要指根据文本的内容提取出该文本所描述的实体对象和对其的观点,并且将结果通过图示直观地向用户展示出来。文本态度倾向性分析算法用于帮助用户了解网络文本中所包含的情感色彩。举例如下:

我们提出新的通过人工标注和机器学习相结合的方法,补充完善现有的情感倾向语言资源,包括词汇、短语、句子、文档等粒度的标注语料的规模,对所包含情感倾向词语的情感强度标注深入优化,并进一步探讨修辞等更深层次的情感倾向获取问题。可以考虑引入半监督的方法来提高情感语言资源的获取效率。

研究结合上下文环境和语义消歧的情感分析,从而更准确地分析词汇在动态的上下文环境中的褒贬含义,比如词“高”在“性价比高”、“价格太高”等环境下有不同的褒贬含义。

将机器学习的文本分类技术与现有的褒贬语言资源以及基于褒

贬语言资源的评分方法更好地结合，从而进行更为精确的情感分析。

结合机器学习和规则方法探讨更好的观点抽取方法，准确抽取语料中的评价实体，以提高针对实体评价特征的褒贬评分效果，比如针对人名、公司、产品等实体的评价。

2.6　关联分析技术

挖掘关联和相关是数据挖掘领域的热点问题，但是目前研究成果的深度应用还不是很成熟，通常的挖掘分析只涉及相关性分析，即只能给出变量之间的相关关系，而更深层的因果关系，一般的数据挖掘方法无能为力。显然，在互联网舆情信息传播相关活动中，发现和挖掘因果关系都是非常重要的，它可以帮助我们对研究对象更本质的理解以及获得一些可以指导行动的知识。如能充分利用数学理论中因果关系的关联分析的方法，通过综合分析某些公共突发事件（比如群体性事件）发生前后网络舆情的变化规律，确定影响事件的关键因果因素，一方面能对该类突发事件进行预警，消除或者降低此类事件的社会影响；另一方面也为应对突发事件制定应对措施提供依据和支持。

1. 将关联分析用于舆情

一般地，将关联分析用于舆情可以按照以下步骤进行。

（1）确定关联分析的对象。

目前针对网络舆情的分析有很多种类，分析对象的选择需要

有根据地进行，寻求具有内在实质关联的对象，如果选择仅仅是表面和虚假的联系，则研究的意义不大。需要在研究之前将所选题涉及的研究对象进行必要的定性和定量分析。

（2）各类舆情发展和网络信息的关系模型。

需要给出实际验证有效的模型来检测网络信息与社会舆情的关联关系，具体可深入到垂直领域舆情内部进行，需要挖掘推动舆情发展的内在规定性。

（3）对于舆情的趋势预测。

网络信息的相关分析用于指导实践和预测舆情未来发展趋势，建立网络信息的回归方程是相关分析的重要环节，利用回归方程在明确各变量的情况下，准确对舆情发展状况进行预测，如能从经验数据中分析出影响舆情信息增长和分布的因素，并建立精确的模型，就能够对舆情信息进行很好的监控。

网络话题传播动态分析的目标是利用关联分析技术分析博客、论坛、新闻等，实现对某个主题的传播趋势进行分析，用动态传播图的形式展现舆情传播的线索。设置舆情传播动态模块对同一主题的论坛帖文、博客文章、网站新闻，进行基于时间的罚分策略计算关联程度分析，以传播网的形式给出同一主题在不同媒介之间的传播关系，结合关注程度分析得出主题的转移趋势，并以平面图、动画以及抽象的有向图形式将示意图展现给用户。网络话题的动态演化分析是通过三维图形下的信息挖掘、叠加检索模型，通过概念挖掘的手段，以波谱图的方式，展现一定时间周期内的舆情

变化情况以及舆情重点和相关关系。系统通过粗细、亮暗、分叉的方式来表达同一时期的报道信息数量、关注度、趋势等，为舆情变化判断提供一定的参考。

2. 网络舆情与社会事件之间的关联分析研究

网络舆情和社会事件之间的关联分析研究包含以下内容：

（1）某些群体突发事件的参与者，或者该问题的关注者在事件暴发前会通过网络媒介交流、组织、扩散相关信息。研究与特定事件相关联的网络舆情因素。

（2）确定这些因素的强度和分布，建立利用这些因素对突发事件进行预警的模型。

（3）利用因果分析模型，在众多网络舆情因素中，确定影响突发事件发生、发展的关键因果因素。从而为应对突发事件的措施制订提供依据和支持。

将分析整理后的信息直接为用户或为用户辅助编辑提供信息服务，如自动生成舆情信息简报、追踪已发现的舆论焦点并形成趋势分析，用于辅助各级领导的决策支持。

2.7　主题检测和追踪

主题检测与追踪（topic detection and tracking，TDT）是舆情分析的重要技术手段，同时也是近十年自然语言处理和信息检索领域的热点研究课题。其主要任务是从连续的记录（如新闻、论坛

发帖、微博等)中识别出系统未知的主题以及与该主题相关的记录,或发现与某一个已知主题有关的新记录。这里的主题可以根据舆情监控的需要来设定。

该研究始于1996年,自1998年以来,国际上每年举行一届TDT评测活动。该评测由美国国防部高级研究规划署(DARPA)和国家标准技术局(NIST)发起,参与者包括DARPA等政府机构,CMU、Cambridge等一流大学以及IBM、GE等公司。该评测极大地促进了TDT技术的发展,取得了大量的重要研究成果。

目前TDT关注的研究重点是事件的检测与追踪,其中,主题是比事件更加宽泛的概念,一个主题可以包含多个相关事件。从本质上看,事件检测是对新闻报道流依据不同的事件做聚类,需要将讨论一个事件的报道归为一类。与通常的文本聚类相比,事件检测的特殊性主要表现在两个方面。首先,事件检测的处理对象是按时间顺序依次出现的新闻报道流,随时间动态变化,而不是静态的封闭文本集合;其次,事件检测是依据报道讨论的事件而不是主题类别进行聚类,所依据的信息粒度相对要小,所以由事件检测得到的类应当更多一些。

在事件检测过程中,主要步骤如下:一是从数据源读入一篇报道,包括内容、时间以及其他相关信息;数据源可能存在多个,报道之间可能没有明显的界限,需要进行报道间的切分等预处理。二是采用质心比较或者最近邻比较策略,计算报道与事件或者报道与报道间的相似度,确定与当前报道最相近的事件。三是若报

道被归入某个事件，则调整该事件；若报道无法归入现有事件，则将其列为新检测到的事件。四是输出检测到的事件，将事件中权重最高的几个特征词或者具有代表性的某个报道标题作为事件描述。

主题检测追踪根据不同用户设定不同主题策略，将互联网作为一个大的语料进行处理，追踪系统能够通过给出某个话题的一则或多则报道，将后输入进来的互联网报道与该话题联系起来，实际实现过程分为两步进行：给定一组样本信息，通过训练得到指定的话题模型；在后续信息中发现涉及目标话题的信息。

由于主题识别与追踪的处理对象是随时间动态变化的语言信息流，不是静态的、封闭的文本集合，因此还考虑了主题追踪的时序性。

整个过程由建立主题模型、基于模型的追踪和产生追踪结果三部分组成。

1. 建立主题模型

判断舆情信息是否与主题相关，需要解决主题的表示模型问题，这里采用向量空间模型来表示，基本思想是：给定一个自然语言文档 $D=D(t1,w1;t2,w2;;tn,wn)$，其中 ti 是从文档 D 中选出的特征项，wi 是该项的权重，$1\leqslant i\leqslant n$。把 tn 看成是一个 n 维的坐标系，而 $w1,w2,\cdots,wn$ 为相应的坐标值，因而 $D(w1,w2,\cdots,wn)$ 被看成是 n 维空间中的一个向量，文档表示为文档空间的向量，以词作为文本特征项。特征词加权采用 tf×idf 加权策略。对

于每个主题来说，需要通过训练信息模型来建立已知主题模型。这时需要计算信息和主题的相似度，采用对称 Okapi 公式算法计算，所得结果为文档和主题之间的分数。

2. 模型追踪

可通过针对关键词的相关计算过程来实现追踪，通过人工提供反馈找出讨论主题的消息，得到每个关键字的权值 λ_{i_n}，然后通过选择打分算法来计算分数。在主题集合(T)中，每个关键字 i_n 的相对概率计算如下：

$$p(i_n \mid T) = \frac{\mathrm{tf}(i, T)}{\sum_{j_n \in T} \mathrm{tf}(j_n, T)}$$

相关主题集合 T、无关主题集合 T_0 和所有主题集合 T_a 的计算都是采用上述进行的，其中 $\mathrm{tf}(i_n, X)$表示关键字 i_n 在集合 X 中的条件概率。

接下来通过手工设置关键字 i_n 的权重 λ_{i_n}，用以调整关键字的条件概率，并将其标准化

$$p(i_n \mid T) = \lambda_i p(i_n \mid T) + (1 - \lambda_i) p(i_n \mid T_a)$$

$$p(i_n \mid T)_{\mathrm{normal}} = \frac{\mathrm{tf}(i, T)_{tz}}{\sum_{j_n \in T} \mathrm{tf}(j_n, T)_{tz}}$$

同样进行在 T_0 中的计算。可通过下面的计算得到相似比值 # rol

$$\# \mathrm{rol} = \lg \frac{p(i, T)_{\mathrm{normal}}}{p(i, T_0)_{\mathrm{normal}}}$$

最后在主题 T 中消息 S 的得分为

$$\mathrm{tfs}(i_n, S) = \mathrm{tf}(i_n, S) \quad \lambda_{i_n} \Big/ \sum_{j_n \in T_a} \mathrm{tf}(j_n, T_a) \lambda_{j_n}$$

$$\mathrm{score}(S, \mathrm{T}) = \mathrm{ok}(\mathrm{d}, \mathrm{cl}) + 1 \Big/ \left(n_S \sum_{i_n \in d} \# \mathrm{rol} * \mathrm{Tfs}(i_n, S) \right)$$

式中 i_n 是出现在消息 S 中的一个关键字；n_S 是消息 S 中不同关键字的数量。通过增量学习的机制，增强当前主题模型的作用，通过合并新的训练信息到主题模型中得到实现。

3. 追踪结果

在上述算法的基础上，通过关键字调整和增量学习的机制，对跟踪结果进行重新追踪，将结果与阈值进行比较，反复多次，得到更可靠的追踪结果。

目前的国内外针对事件检测的研究存在着一定的改进空间，主要体现在：

(1) 热点事件排序问题，人们往往没有时间去查看大量的新闻事件，所以最热的新闻事件排序应该越靠前，这就涉及排序算法的设计。

(2) 事件相似性问题，由于对同一个新闻事件不同方面进行报道的新闻可能相似度较小，从而使得同一个新闻事件在事件发生初期被分为多个小事件，进而随着事态的不断发展，这些事件的相似度可能会越来越大，这样就可能给用户的浏览带来迷惑和不便。如能给定适应该问题的科学的相似性度量方法，便可以对相同的内容进行剔除。

(3) 新闻报道淘汰问题，在实际应用环境中，事件检测是一个

长期持续的过程。随着事件的动态演化，事件内的一些新闻和该事件的相关性在逐渐降低。另外，周期较长的事件随着时间的积累也可能出现膨胀现象，整个事件内容过于宽泛。

(4) 事件描述问题，目前新闻事件的描述有两种方法。选取该事件中最重要的若干个特征词，或者该事件中某个新闻标题。由于中文自然语言处理技术还不够成熟，提取的特征词往往很难完整准确地描述事件。而如果用事件中某个报道标题作为描述，对于一些综合性的事件，则该报道可能仅是事件的一个方面，对事件的描述不够全面。

未来的研究应该着重于利用新闻事件本身的特点，解决热点事件排序、事件合并与调整、新闻报道淘汰，以及新闻事件描述等问题，实现对持续新闻流进行动态、高效的事件检测和热点追踪。

目前话题检测方法在新闻等较正规语料时效果较好，应用于论坛（很多短文本）、博客（很多特别长的文本）等不规则数据效果尚需进一步改进，需要对非正规文本，例如短文本和长文本（包含很多内容）进行特殊处理，以保证较好的检测效果。

2.8 舆情热点发现和监测

互联网舆情信息量巨大，如能构建针对舆情的排名和预警系统，对当前舆情热点进行重点关注，将会大大提升监控效率与监管效率。针对舆情热点发现技术目前存在着的一些缺点和不足，以

下几点是可以重点关注的：

1. 关键词的选取，对不同频率词汇变化的权重度量

基于统计的方法进行关键词的提取，同时需要结合词典技术，可充分发挥基于统计的算法不受句型限制的优点，又可利用关键词词典控制统计算法中的噪音项。最后提取关键词的结果作用于词典以增强词典的权威和可信度。

2. 词频显著性波动的预测，使用历史均值和历史标准差进行

通过累积舆情信息数据，结合统计学方法，考虑统计关键词单位时间内词频的历史均值与历史标准差，针对关键词词频的波动来发掘事件运行的趋势，可以通过一些数据挖掘的方法，如支持向量机来进行预测。

3. 对高频词和低频词类别的区分及事件异常度的检测

定义关键词的权重，用以关键词的分类，主要通过利用文档中提取的关键词和相应统计信息，构建分类模型来区分文档中关键词的类别，进而利用历史均值和权重计算关键词的热度，通过设置相应的阈值检测关键词对应出现的事件异常。

4. 热点事件预警及显示处理

设计并实现热点事件关键词及其热度的波动图形区分显示，直观地表达事件异常，并提供用户预警功能。

对于信息热点的发现和检测部分的研究，可参照已有的针对金融市场信息异常波动的研究成果来进行，比较方便构建相应的模型和算法实现。

5. 异常度阈值的判断

建立在对事件相关词汇的异常度进行计算的前提基础上，阈值的判断除了需要引入历史标准差之外，还需要区分词汇的频度高低，不仅仅是绝对数值的问题，需要具体化到实际应用的场合。在已有金融市场信息的异常发现研究基础上，已经建立了异常阈值的设定和判断标准，需要通过进一步的扩展处理延伸到其他领域的信息异常判断中。

随着网络日益成为人们发布信息、沟通信息的主要媒体，网络上的信息也越来越能反映人们关注的焦点和社会热点事件了。因此，监控网络信息中所反映的热点问题和热点事件就成为自然的需求了。不论是普通用户还是行业专家都希望实时地跟踪他们所关注领域的最新热点话题或者新闻，以便了解该领域的最新进展。

不难发现，在一般情况下，互联网信息中某个关键词大量集中出现往往意味着某个热点新闻或者热点事件的发生，而当发生了被广泛关注的新闻或事件时，又会在网络上集中出现大量的带有相关关键词的文本。因此，互联网文本中热点关键词数量的较大变化常常反映了社会热点新闻或事件的出现或降温，而网络上反映热点新闻或事件的文本又会进一步推动广大网民对于相关新闻和事件的关注程度和看法。也就是说，异常高的关键词词频与显著的热点新闻和事件有一定的耦合关系。所以，这里，为对舆情信息进行研究，应避开对词频小的变化问题的预测，只关心异常高的词频变化量。这个技术点对于网络监管机构、关注社会热点新闻

和事件的机构来说，是非常有价值的自动跟踪热点词汇出现频率的技术。基于上述观察，对舆情热点发现和监测，我们讨论一种基于UDAHIN（Universal Detection Algorithm for Hot Internet News）技术的舆情信息热点排名方法。

不同词有不同的出现词频，而在某日，不同出现词频的词的相同出现次数有不同含义。对于一个使用频率很高的词来说，词频的历史均值和历史标准差都很大，例如，分别是 500 次/天和 350 次/天。如果在某一天，其互联网频率增加了 300 次，变成了 800 次，即增加了大约 1 倍，那么一般仍然很正常，但是，如果其互联网频率变成了 1200 次，即增加了大约 2 倍，就预示着发生了相应的热点新闻或事件了。

而对一个频率比较低的词，平均日互联网出现频率及其标准差很小，例如，分别是 20 次和 15 次。如果在某一天，其互联网频率增加了 30 次，变成了 50 次，即增加了大约 1 倍多，那么一般仍然很正常，但是，如果在某一天，互联网上该词的信息量增加了 300 次，变成了 320 次，则预示出现了相应的热点事件或新闻。

也就是说，同样是增加 300 次，对高频词说，仍然正常；而对低频词来说，则说明出现了异常事件。即对具有不同词频的词的度量，标准是不同的。对于低频词，上述的 300 次出现次数称为异常高的词频增加量。这里在热点信息排名上的主要目标是监测异常高的词频增加量，进而预测网络热点信息的出现或降温，并进行必

要的报警。

Khoo 等人于 2001 年提出了一种跟踪热点话题的方法，对一些定点的网站或者网页定期统计一些关键词项（term）的词频，并利用 tf×idf 公式计算每个 term 的当前权重，并从中得到当前的热点话题。其贡献之处在于，给出了一种标准化的公式来计算每个term 的当前权重，随着时间的变化，这个权重也会随之变化，从而反映出互联网信息热点的变化情况。其主要缺点在于，没有考虑每个 term 的历史均值和历史标准差，因此无法按照高频词和低频词的历史表现对异常的热点进行准确的度量，只能对各个 term 进行横向的比较。

2.9 本章小结

舆情分析的技术有很多，典型的包括关键词提取、摘要提取、文本倾向性分析、关联分析技术、主题检测和追踪、舆情热点发现和监测等。在关键词提取中，tf×idf 是较为普遍采用的方法。文档摘要技术在图书馆领域和自然语言处理领域有很多应用。文本倾向性分析可以得出文本舆情所描述的情感，在一定程度上代表了作者的感情取向，决定了文本内容的褒贬含义。主题检测追踪及舆情热点是舆情分析的重要技术手段，它将会提升舆情的监控与监管效率。

思考题

1. 简述社会互联网舆情分析的主要技术。

2. 掌握关键词提取及摘要提取技术。

3. 总结文本倾向性分析的主要步骤。

4. 了解网络话题传播动态分析、因果分析模型及舆情热点发现和监测方法。

5. 试列举网络话题传播动态分析的应用。

第3章

社会网络中的用户行为

本章学习目标

- 熟悉基于用户行为的社区网络概念。
- 了解社会网络中的“社交圈”与“兴趣圈”。
- 掌握社会网络用户行为的主要方法。

3.1 引言

区别于一般的互联网，在社会网络中，有一大类是社交网络。这类网络的用户，不但有内容，用户间还形成了网状结构（以微博为例，如图 3-1 所示）。

用户去“粉”哪些不同的用户，以及“粉”多少其他用户，这些行

为，直接造成了用户在社会网络中所处的位置（包括所在社区、在什么样的路径以及用户的重要性）。

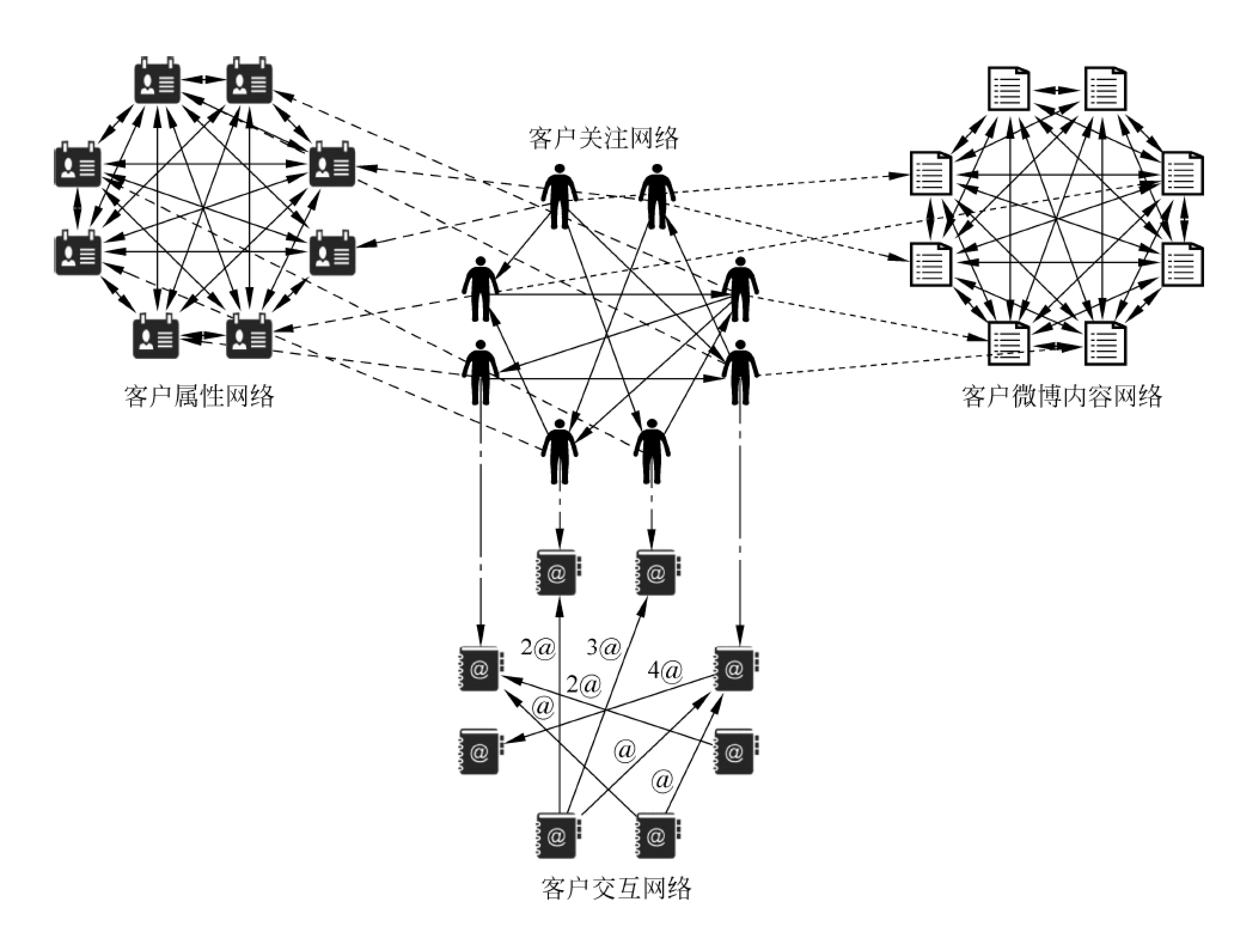


图 3-1　微博平台上用户之间的关系网络

在不同的社会网络中，用户之间往往存在某些共同特性，即网络的群体特性。一般情况下，把内部联系紧密、外部联系稀疏的一群用户称为社区，它反映了网络元素之间的拓扑关系和功能实体，在不同的应用领域，社区代表不同的实体关系群。从巨大的社会网络中挖掘出社区的过程称为社区发现，是社会网络分析的一个基本任务。

3.2 基于用户行为的社区网络

社区发现研究已经受到研究者们的广泛关注，社会网络通常采用图结构表示，针对无向无权值的网络，Newman 和 Girvan (2004)提出了经典的 GN 算法，边界数概念是 GN 算法的核心内容，通过删除边界数高的边从而分裂得到整个网络的社区结构，但是，GN 算法由于计算量大而很难适用于用户数目上万级的大型复杂网络。同年 Newman 在 PANS 会议上又提出了衡量社区发现优劣的模块度 Q 概念，之后，最大化模块度的自下而上的合并算法和优化模块度的发现算法被研究者广泛提出。尽管最大化模块度 Q 一度成为衡量社区发现优劣的依据，但是由于其依赖于全局的网络拓扑结构，会导致大的计算量，而且，分辨率限制的问题也是模块度优化方法的症结。后来有的学者提出了派系过滤算法(clique percolation method，CPM)，CPM 算法的本质是认为典型的社区应是全连通的完全子图，全连通子图之间共享的用户是重叠用户，其主要目的就是找到紧密相连的完全子图，尽管 CPM 算法对社区的发现一般来说是很有效的，但搜索完全子图是非常耗时的，而且完全子图的大小值 k 不易确定。后来有的学者提出的 LFM 算法从局部拟合构造社区的发现算法，局部性反映了社区的自然特性，但随机选择初始用户会影响 LFM 算法的社区发现结果，它与以往从点的角度划分社区的思路不同，其核心思想是将边

看做点,重新构造点之间的相似关系而构造新的图结构。图的另一种有效的表示形式是邻接矩阵,因此基于邻接矩阵的谱分类方法也是社区发现的常用方法。谱分类方法的核心是构造邻接矩阵的拉普拉斯矩阵,通过拉普拉斯矩阵的第二小特征根来判断用户所属的类别,当网络的确是近似地分成两个社区时,用谱平分法可以得到非常好的效果,但是,当网络不满足这个条件时则不行。

随着网络不断变化,社会网络的形式表现出多样化。多样化表现突出的一个方面是网络链接有方向性,在实际的复杂网络中,链接关系时常表现出非对称性,比如 Twitter 的用户关注关系、科技文献网络的引用关系、网页之间的超链接关系等应用网络。因此,人们依据信息在无向网络中的传播行为特性,以及有向网络中信息游走的方向性,提出了共社区的邻近相似性测量方法,衡量用户在同一社区的通路相似性,应用邻近相似性可以有效地将有向图中的方向转化为方向权值,从而合理地将有向转化为无向网络。

在用户行为角度看来,社会网络能够增强用户体验,从初始的“需求识别”和“产品发现”,到“产品选择”和“产品参照”。

(1) 产品发现行为:在顾客购物行为的初始进行需求认可和产品搜索的阶段,社会网络能够帮助顾客预估一些新的产品。例如,在网上论坛和社区里讨论能够非常有用地帮助顾客更加明晰内心的需求,同时社会网络中的一些顾客推荐或参照的有感染力的内容往往能够帮助用户与那些了解或值得信任的朋友一起讨

论、发掘新的产品。从商业角度看，社会网络在产品发现方面的应用印证了其“意识助推器”的角色。

（2）产品选择行为：在顾客进行实际产品选择行为的阶段，社会网络利用一些独立的第三方评测或专家建议来推动顾客的购买决策。譬如，便携式社会化图谱和顾客评级及评论软件。社会网络在产品选择方面的应用扮演了“决策催化剂”的角色。

（3）产品推荐行为：购买后，社会网络能够帮助人们分享其购物经历，将一些感受反馈给其他人。例如，网上的口碑记录应用能够帮助商家们意识到顾客参照推荐的价值——构建客户忠诚度。社会网络在产品推荐方面的应用扮演了“主张活化剂”的角色。

这里的每一步都彰显着社会网络的重要性，然而企业是否真的理解用户行为却存在着不确定性，如推荐的产品是不是对，新产品是不是反响不错，产品是不是存在硬伤，这些都需要舆情来给出答案，社会网络正是天生的舆情产生地，科学的舆情研究将确认企业的正确策略，坚持给客户更好的体验，同时，通过舆情发现不足乃至重大错误，及时纠正，挽回损失，更好地服务用户。

由于以社交网站、微博、微信、视频分享网站为主的社会网络的营销尚不清晰，因而，找出社会网络的整体特性以及社会网络中客户行为的共有属性，可以帮助我们更好地思考社会网络的本质。

3.3　社会网络中的“社交圈”与“兴趣圈”

“社交圈”与“兴趣圈”是促进社会网络持续繁荣、维持客户关系稳定、保持参与者忠诚度的重要参考准则。所以，对于社会网络用户的属性，我们可以从“社交圈”与“兴趣圈”两个层面来分析。

1．社交圈

从狭义上来说，社交圈是我们日常生活中与朋友、同学、同事之间的各种关系网络构成的一个人际圈子，从广义上来说，社交圈可以延伸为我们每个人的生活圈。

社交圈的存在让营销者更为欣慰。以人人网、Facebook为例，大多数的注册会员都是来自于各个高校的学生，那么即便日后会面临毕业、工作、生活等变迁，但是这些稳固的同学关系却永远不会消失。同样，在以白领为主要客户群体的开心网上，同事间的关系虽然相对于同学来说因为变化更大而淡薄一些，但是依然是一种相对稳固的人际关系（Poyry，2013）。互联网社交圈的建立与发展还取决于社交网站客户的交互方式（Backstrom，2006）。社交圈中不同身份的人的影响力也不尽相同（Stutzman，2006）。

社交圈对于网络营销来说无疑有着积极的促进作用，无论是品牌的创建或者是促销的推进，在稳定网络中的传播广度与深度，都会比大众网络更有优势。

社交圈的分享也可以分为两种：一种是消费者主动的分享，

一种是营销者促使的被动分享。被动分享一般都是由网络营销者来促成的，发起者往往会针对消费者和潜在客户的商品促销信息与订单信息的分享、发布行为给予一定的奖励，以这种激励方式达到促销以及品牌传播的目的。

2. 兴趣圈

从理论上说，兴趣圈的范畴应该要比社交圈更大一些。相对于社交网站的社交圈为主的状况，微博、群组、知识分享平台、视频图片网站、团购、LBS 服务等，大多都是基于兴趣形成的社会化的关系网络。当然，广义上来说这种因为兴趣而产生的关系网络也属于社交范畴，但是单纯从社会网络的类别、数量上来区分的话，因为兴趣而形成的社会网络种类更多、品种更齐全。

在常见的社会网络中，除了社交网站之外，还存在大量的其他网站。比如以维基百科、百度百科为代表的知识分享平台；以 Twitter、新浪微博、腾讯微博为代表的微博平台；以 YouTube、优酷为代表的视频分享网站；以 Flickr 为代表的视频分享网站；以豆瓣、百度文库为代表的文档分享网站；以大众点评为代表的消费评论网站，甚至包括各种团购网站以及以微信为代表的移动客户端上的平台，其实都属于兴趣圈网络的范畴。

以兴趣圈形成的网络具备一些特殊的特性，那就是社会网络的自优化特性。以最典型的新浪微博为例，对于一个商家或者公众意见领袖来说，普通的网民与之形成的关注但是不相互关注的关系从本质上来说并非是社交关系。同样地，这种大多数时候的

单向互动也完全不符合社会网络的互动特性，而只是类似于更多的媒体中心，进行单向的信息传递而已（Bouras，2004）。

对于商家来说，交互从来也都不是充分的，而这正是社交圈之所以存在的原因和价值。社交圈可以分为陌生人购物分享网络，例如蘑菇街、美丽说，以及熟人圈社交网络例如微博、人人网等。对于商家来说，他们更关注的是这些人是否真正对他的商家和产品感兴趣，而对于每一个客户的深度交流，只能作为一个远景却未必能实现。社会网络活动中的自优化功能会帮助商家找寻到最忠实的客户，也就是长期留存下来的稳定网络。

3.4　社会网络用户的行为

社会网络的用户有其自身行为规律。

首先，研究发现（曹润，2012），从数量上看，原创型微博与转发型微博几乎各自占据了“半壁江山”，原创微博所占比例略高于转发微博所占比例，这说明发表观点与传递信息在微博中有着几乎同等的重要地位。原创和转载占总体的 92.21%，占绝大多数，且转载数远远大于评论的数量，这与微博中转发过程常常伴随着评论的特性有关，如图 3-2 和图 3-3 所示。大 V 用户是最为“勤劳”的内容贡献用户，平均微博字数为 53.65 个字，远远高于其他类型的用户，其原创微博的平均字数高达 81.71 个字，可以看做大 V 用户的原创微博中包含更大的信息量。女性用户原创微博字数比男性

高很多，且女性用户原创微博的字数是转发的近 4 倍，而男性用户原创微博的字数是转发微博的字数的 3.45 倍，显示性别对“微博风格”会产生比较大的影响。例如女性可能更倾向于详细地描述一个事件或一种物品，而对转发的内容简短地发表意见，而男性用户倾向于相对简短地叙述。

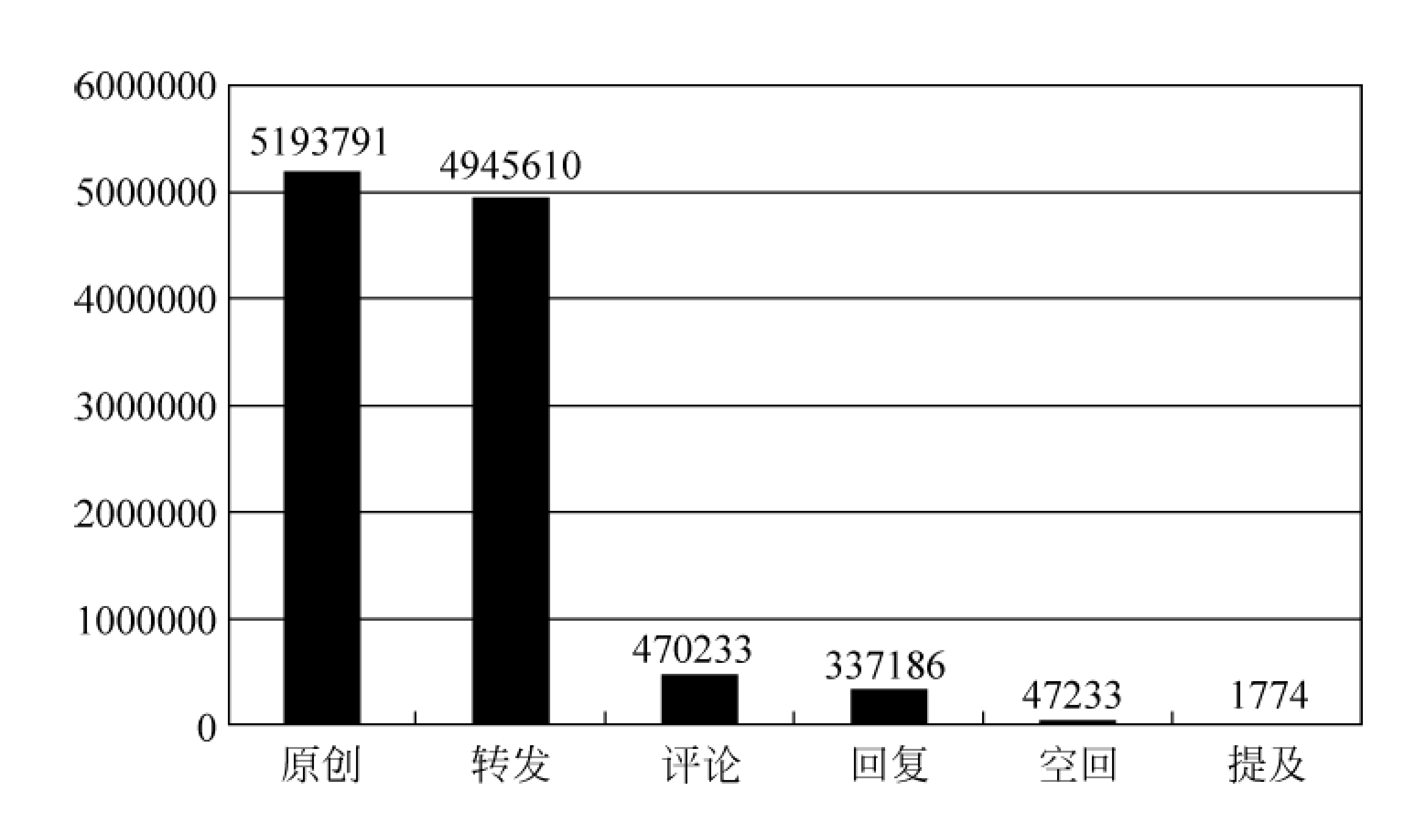


图 3-2　不同类型微博数量分布

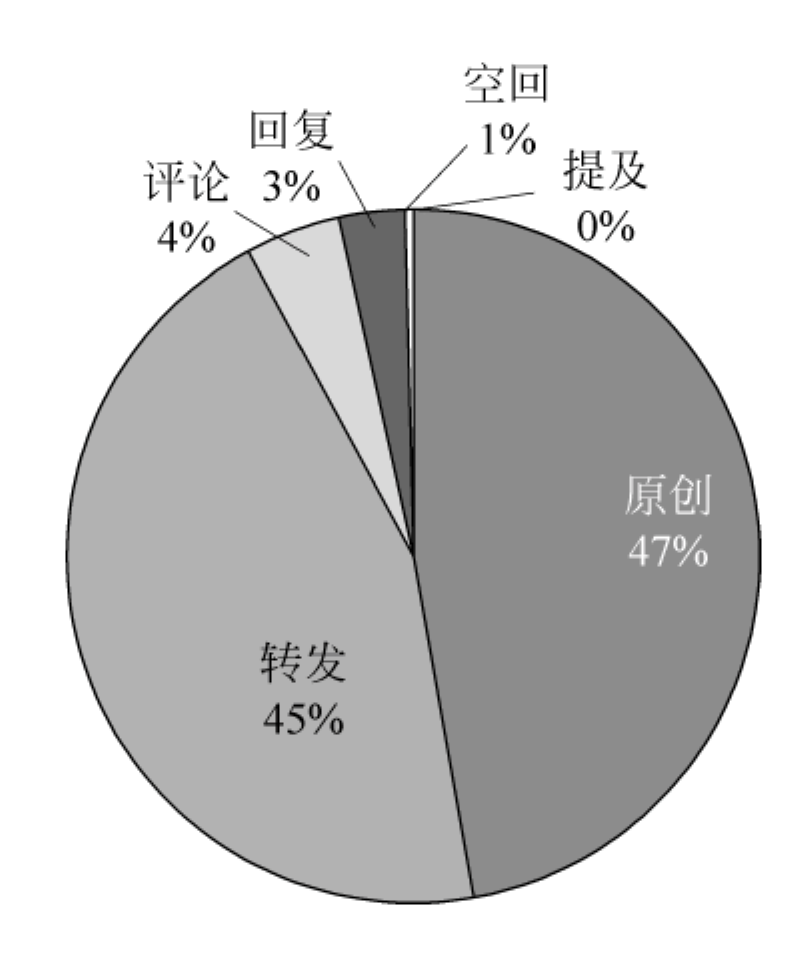


图 3-3　不同类型微博比例分布

其次，不同类型的微博转发的次数不尽相同如表 3-1 所示，包含图片的微博转发数均值为 256.2，不包含图片的为 26.4。很明显，不包含多媒体信息的微博平均转发量远远低于包含多媒体信息的微博平均转发量，说明包含多媒体的微博携带的信息量更大，带来了更大的转发量，更容易造成较大的影响。

表 3-1　不同微博类型的微博转发数

	包含多媒体	不包含
图片	256.2	26.4
音乐	456.7	66.6
视频	221.3	66.1

再次，微博转发影响力也不尽相同。我们曾完成如下实验。假设没有被转发的微博没有造成转发影响，以转发次数与微博数目的乘积之和作为整体的微博影响力。实验发现，通过分析微博的转发次数，并与其微博用户关联，显示微博转发数 count 值大于 357 的微博转发量累计占 80%，仅来自 1367(7.93%)名用户；count 大于 93 的微博转发量累计占 90%，来自 2124(12.32%)名用户，显示 10%左右的用户贡献了 90%的微博影响力。

我们进一步根据微博类型进行分析，对于原创型微博：count 大于 695 的微博的累计转发量占据总转发量的 80%，来自 965(5.60%)名用户；count 大于 217 的微博的累计转发量占据总转发量 90%，来自 1402(8.13%)名用户。对于转发型微博：count 大于 60 的微博累计转发量占中转发量的 80.1%，来自

1633(9.48%)名用户;count 大于 29 的微博累计转发量占总转发量的 90.1%,来自 2009(11.66%)名用户。原创聚集现象更加明显,更加少量的活跃用户贡献了更多的内容。

我们的研究选取腾讯微博 17232 名用户,共计 10995827 条微博,分析研究微博的用户内容生成模式。研究发现每日的发布微博数曲线呈现波动上升的趋势;较少量的用户发布了绝大多数的微博;不同类型的用户有着不同的“微博风格”。

我们总结出微博用户内容生成的一些特点:

(1)“90-10”规则:微博显示出了一种用户内容生成更明显的聚集性,表现出一种“90-10”规律,更加少量的活跃用户贡献了更多的内容,如 13.19%的用户发布了 90%的微博;更加集中的某一部分微博造成了更大的影响力,原创型微博的总转发影响力的 90%来自 8.13%名用户,转发型微博 11.66%的用户贡献了 90%的微博转发影响力。

(2)用户所发原创微博所含字数均明显大于转发微博,在女性用户中的差距更大,大 V 用户所发微博平均长度是非大 V 用户所发微博长度的 1.5 倍。

(3)包含多媒体的微博携带的信息量更大,带来了更大的转发量,更容易造成较大的影响。

用户内容生成分析在基于微博的研究中有极其重要的价值,同时随着微博用户大规模增加,微博内容的影响也不断扩大,微博的用户内容生成分析对了解信息分享型的社会网络的本质规律有

着极大的意义。在今后的工作中，我们会收集更加全面更具代表性的数据，对数据进行更深层次的分析，挖掘出更有效的用户行为模式，以优化相关研究结果。

事实上，网络用户行为的研究与心理学、社会学、社会心理学、人类学以及一切与网络行为有关的学科密切相关，它研究网络用户行为的规律性，借以控制并预测网络用户行为，并为实现政治的、经济的和文化的目的服务。具体讲，网络用户行为研究就是分析网络用户的构成、特点及其行为活动上所表现出来的规律。

目前有关社会网络用户行为的分析研究，主要集中在以下几个方面：

（1）基于用户行为模型的方法。王实等（2012）提出了一种基于隐马尔科夫模型的兴趣迁移模式发现方法，通过分析用户迁移模式间的关联规则来发现用户行为特征和兴趣迁移变化；张振国等（2013）提出了基于序列模式挖掘的社交网络用户行为分析模型，该模型以社会网络群体用户为研究对象，引入序列模式挖掘的方法实现用户行为分析，最终获得用户的频繁行为序列模式，其分析结果能够反映用户行为之间的有序相关性，因此可以为策略制定提供更好的支持；肖玉芝等（2014）提出了利用超图的数学理论建立用户行为的超网络模型，通过分析实体用户、用户活动、用户兴趣三维度的映射关系，在某论坛的真实数据验证了该模型能够快速定位用户并刻画出用户兴趣爱好的差异性。

（2）基于统计方法分析用户行为的方法。何静等(2013)提出了基于统计学的方法从个体和群体层面对微博的网络拓扑结构和用户的行为进行统计分析，研究发现微博用户的行为表现出多重的标度特性和复杂性，在此基础上，结合人类行为动力学理论得出了微博用户行为的一些共性。

（3）基于用户行为特征的方法。Anagnostopoulos 等(2008)通过研究社会影响力，找出了社交网络用户行为的变化原因，并在此基础上分析用户行为变化趋势；Goyal 等(2010)通过研究，解决如何从用户的历史记录里学习影响的概率，并得到分析预测结果；史学敏通过研究时区差异对社交网络用户行为的影响(史学敏，2011)，发现社交网络用户行为呈现时区特性，以此建立了用户行为特征。

3.5 本章小结

对社会网络的研究，不仅要讨论其节点的内容，还要探索其节点之间“织成”的网络结构。这些结构上的关系，直接造成了节点用户的行为。网络结构形成了各种各样的“社交圈”与“兴趣圈”。有关社会网络用户行为的分析研究，主要包括基于用户行为模型的方法、基于统计方法分析用户行为的方法以及基于用户行为特征的方法。

思考题

1. 了解基于用户行为的社区网络。
2. 简述社会网络中的“社交圈”与“兴趣圈”的内容。
3. 试列社会网络用户行为的分析方法。

第 4 章

企业网络舆情管理的模型

本章学习目标

- 熟悉企业在线舆情的分析预警管理模型。
- 了解企业在线舆情的干预处置管理模型。

4.1 引言

社会网络大数据环境下的企业舆情也为研究舆情对企业经营管理绩效的影响、舆情传播发展规律以及舆情管理方法手段的效果提供了良好的条件。作为数字化记录的用户内容，在线企业舆情具有良好的可追溯性，使得企业舆情的定量刻画成为可能。比如，网络用户历史上发布的关于企业的文字、图片、多媒体等信息

均可在网络中检索得到，从而汇总得到企业舆情的历史记录。将企业舆情的传播发展走势与企业经营管理绩效和股票市场表现进行分析，便可得到舆情对企业经营管理绩效的影响；具体分析企业舆情传播、发展的历史变化情况，可以得到企业舆情传播发展的一般规律；更进一步的，通过对比观察企业采取舆情管理方法手段后企业舆情的发展情况，可以得到对应舆情管理方法手段对企业舆情传播发展的作用。例如，电子商务平台上在线评论的出现和积累，使得营销领域的学者得以实证检验消费者口碑（word of mouth）对于企业产品销售的影响（Chevalier 和 Mayzlin，2006）。例如郭小钗和陈蓓蕾（2009）探讨了网络上的口碑效应，并研究了这些效应与购买意愿的关系。郑小平（2008）从商家视角讨论了在线评论与购买决策的关系。施晓菁和梁循（2015）讨论了在线评级和评论对消费者购买决策的影响。

同样，随着社会性应用的广泛使用，用户间的社会网络关系及信息交换过程也得到了数字化的体现。比如，对于网络用户，可将其表示为社会网络中的节点；对于网络用户间加好友、设关注等操作，可将其表示为社会网络中节点间关系的建立；对于网络用户阅读好友博文及转发、分享好友日志等行为，可将其表示为企业信息在社会网络内的传播。这就使得企业舆情研究可以考虑其中用户间的社会网络背景：深入分析网络中不同类型节点在企业舆情传播、发展中的作用，网络结构对于企业舆情传播、发展的影响，并针对特定的网络结构设计相应的舆情管理方法手

段。例如，以往关于公共舆情的研究往往根据假定的社会网络结构及意见影响机制分析给定初始状态的意见收敛结果，通过建模及模拟实验得到相应的结论(Nowak 等，1990)。而在线记录的出现使得以往的模型得以实证检验，并为更加符合实际的新模型的提出提供了可能。例如，在企业产品舆情方面，Ye 等(2011)通过使用多元回归的统计方法，实证研究了企业绩效和网络评论的关系，他们发现，在旅游业，好的评论可以使销售量增加 10%左右。

社会网络舆情对企业经营绩效存在着相互影响的关系如图 4-1 所示。在社会网络舆情的影响下，营销管理、客户关系管理、生产管理、信息管理、财务管理都需要在企业舆情管理的驱动下随之变革和创新。优质的企业舆情管理可以改善企业的客户关心，促进销售，从而给企业带来充裕的资金，提供企业的绩效；反过来，好的企业绩效也会在社会网络上产生正面的企业舆情。

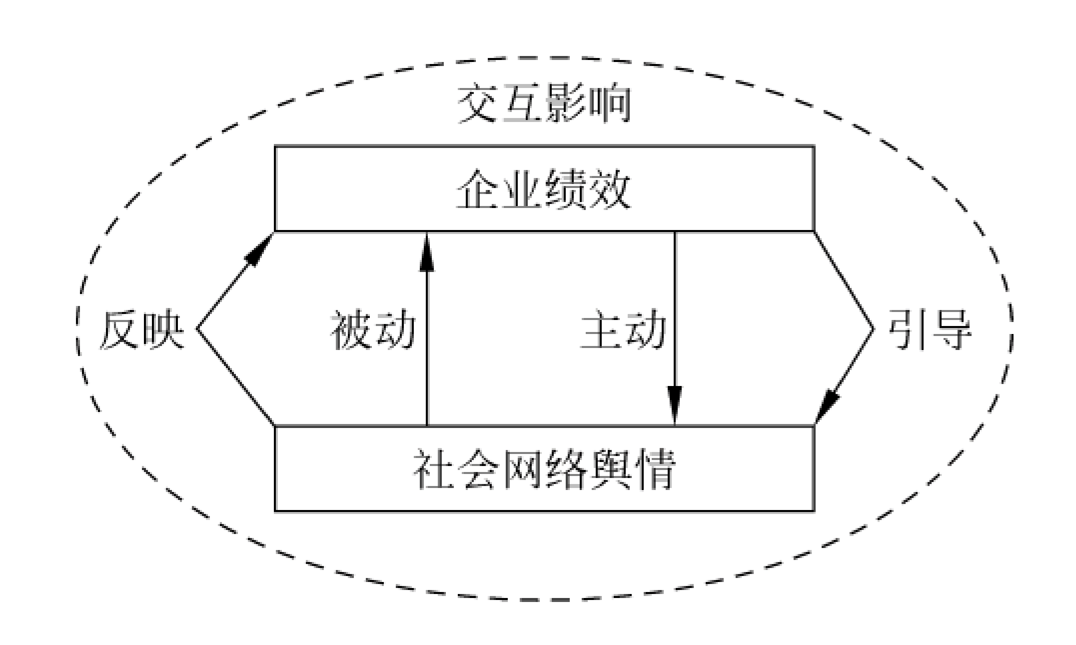


图 4-1　网络舆情对企业经营绩效的相互影响关系

4.2　企业在线舆情的分析预警管理模型

企业网络舆情分析与预警就是根据企业网络舆情态势信息，利用机器自动推理技术，对企业网络舆情的威胁程度进行定量估计，做出企业网络舆情的预警等级预报，并进行可视化。其中，舆情的预警等级可以划分为：轻警情（Ⅳ级，非常态）、中度警情（Ⅲ级，警示级）、重警情（Ⅱ级，危险级）和特重警情（Ⅰ级，极度危险级）四个等级，并依次采用蓝黄橙红四种颜色来加以表示，如图 4-2 所示。

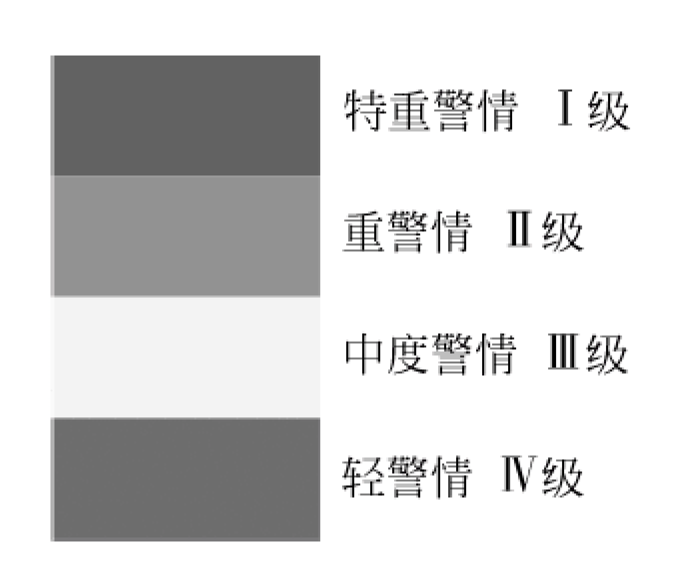


图 4-2　舆情警报分级颜色示意图

- 蓝色级（Ⅳ级）：该类舆情所受的关注度低，传播速度慢，影响范围小，不可能转化为舆论。
- 黄色级（Ⅲ级）：该类舆情所受的关注度较高，传播速度中等，具有一定的影响范围，不可能转化为舆论。
- 橙色级（Ⅱ级）：该类舆情受到很高的关注，传播速度快，影响范围很大，各类媒体都较为关注，有可能转化为舆论。

- 红色级（Ⅰ级）：该类舆情受到极高的关注，传播速度非常快，影响扩大到了整个社会，即将化为舆论。

在已有的研究中，主要包括四类模型：基于模糊推理的网络舆情预警模型、基于情感倾向分析方法的舆情预警模型、基于 Web 数据挖掘方法的舆情预警模型以及基于网络舆情指标分析方法的舆情预警模型。

1. 基于模糊推理法的舆情预警管理模型

基于模糊推理的网络舆情预警方法主要是引入了“战场威胁估计”理论（汤志荔，2011）来进行研究。战场威胁估计指的是对敌方的攻击能力及威胁程度进行定量评估，并将判别结果划分等级。

这种方法利用网络舆情社会学研究成果，分析网络舆情对社会影响的主要因素，选取包括话题重要度、情感倾向性、话题关注度、公众数量、传播速度在内的 5 种网络舆情分析指标；从计算机可实现性出发，对各分析指标及舆情预警等级进行模糊化，构建相应的网络舆情分析指标，在此基础上，采用模糊推理技术自动判断网络舆情预警等级（李弼程，2010；杜智涛，2013；李纲，2014）。

该方法将网络舆情中的话题对应于战场的目标，将网络舆情中的公众对应于战场的环境（黄晓斌，2010）。这种预警方法的主要过程是利用战场威胁估计的思想和方法，分析网络舆情话题，得出其中对社会影响程度较高的属性，利用此属性来自动评估其对社会的影响程度，得到一个评估值，划分预警等级，从而达到进行自动预警的目的。

2. 基于情感倾向分析法的舆情预警管理模型

网络舆情危机预警的成功与否，主要体现在能否每天从海量的网络信息中捕捉到或分析出潜在的重大舆情生成点，而从技术角度进行自然语言处理的情感倾向性分析技术，能从大量的信息中提取信息提供者对于某一对象所持的立场，识别信息中所包含的主观意见（张超，2008；吉祥，2010；项斌，2010；丁菊玲，2011）。

情感倾向性分析也被称为情感分类、情感分析、文本意见挖掘、观点挖掘等，涉及自然语言处理、信息检索、数据挖掘等研究领域。一般分为文档级观点挖掘和语句级观点挖掘，其情感倾向包括简单的赞同、反对、中立三种态度，也包括对某一对象所持态度的强度，甚至舆论对该对象的具体看法和态度等。该方法主要从海量的 Internet 数据中采集舆论的相关信息，获取广大民众的主观观点，并依据民众的观点对可能引发网络舆情危机的事件或舆论进行预警。

3. 基于 Web 数据挖掘法的舆情预警管理模型

基于 Web 数据挖掘的预警就是从网络中提取与目标相关的数据，对网络数据进行网页特征提取、基于内容的网页聚类、网络内容关联规则，得到与 Web 挖掘目的相关的目标数据集，然后通过站点识别、数据选择、数据净化、用户识别和会话识别等数据预处理的方法对目标数据集中“杂质”数据进行过滤，将多个数据源中的数据统一存储，利用路径分析、关联规则挖掘、时序模式发现、聚类和分类等挖掘技术从存储数据中挖掘出有效的、新颖的、潜在

的、有用的及最终可以理解的信息和知识，最后利用合适的工具和技术对挖掘出来的模式进行分析、解释，并能够根据分析结果对网络舆情进行危机预警（张亮，2009；何佳，2010；张一文，2012）。

利用 Web 数据挖掘进行预警有以下四个步骤：

（1）采集数据。即从 Web 中提取与网络舆情事件相关的数据，形成目标数据集。

（2）预处理数据。过滤目标数据集中的数据，将杂质数据去除掉，并将不同数据源的数据统一存储，便于数据挖掘。

（3）模式发现。利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识。

（4）模式的分析和预警。利用合适的工具和技术分析解释挖掘出来的模式，并能够根据分析结果对网络舆情事件的危机进行预警（李季梅，2009）。

4. 基于网络舆情指标分析法的舆情预警管理模型

在网络舆情的发生、发展的过程中往往会表现出一系列的特性，所以很多学者在研究分析了网络舆情的基本特征后，遵循一些原则，采用一定的科学方法来确定网络舆情的关键指标、指标维度和分析指标量化等，建立网络舆情危机预警指标体系，然后运用一定的科学合理的数学方法建立预警模型，进行网络舆情的预警（曾润喜，2009 和 2010；戴媛，2009）。

目前对于网络舆情危机预警的指标体系的研究已经越来越多，但是仍然存在很多不足，主要有：

（1）末级指标量化问题。由于网络舆情很多特征都是通过网络中的语言、图片、声音甚至传播网络结构等体现的，指标量化时，尽管不少文献给出了详尽的末级指标量化说明及数据获取方法，例如李实等(2009)探索了面向博客、BBS 等网站评论内容中产品特征提取及相关技术，但是在实际应用中，往往会很难以获得需要的数据，或者获得数据的效率太低等问题，从而影响危机预警的及时性和有效性。

（2）缺乏防御性预警指标。现在大多数网络舆情危机预警的研究都是用以进行事后评估的指标体系，同时寻找较好的防御性预警指标比较困难，效果无法进行有效评估。

（3）大多指标体系包含较多的定性指标。一般定性指标都是通过专家打分法获得，存在一定的主观性，就可能会影响对网络舆情信息判断的客观性和准确性。

4.3　企业在线舆情的干预处置管理模型

1. 基于生命周期理论的舆情处置管理模型

生命周期(life cycle)是指将事务视为“从摇篮到坟墓”(cradle-to-grave)的生命运动过程的一种表述(谢科范，2010)。生命周期理论认为，事物一般经过发展、成长、成熟、衰退等几个重要阶段，周而复始循环反复。在信息科学领域，信息的内在价值与使用价值会始终跟随客观世界的运动而转移变化，并表现出类似有机生命体的周期性运动特征(林光，2005)。

企业网络舆情事件产生后一般都会经历酝酿期、爆发期、高潮期、消退期和平稳期五个阶段的生命周期（史波，2010；谢科范，2010；王国华，2012；周伟恒，2013）。舆情事件发生后，首先进入酝酿期，到达一定程度后，随即进入爆发期，以较快的速度进行传播，在短时间内便会进入高潮期。在高潮期持续一段时间后，企业网络舆情进入消退期，其消退速度由快而慢，切持续时间较长，企业网络舆情在消退期间还可能出现反复。消退期结束后便进入平稳期，但在平稳期，企业网络舆情影响并不为零，而是起伏比较小、影响力比较小。

明确企业网络舆情的信息生命周期对提升企业网络舆情的管理水平具有重大意义。企业网络舆情的生命周期管理与其相关联领域活动的生命周期相结合，揭示了同一主题领域的网络舆情的发展过程，可以使我们更有效地把握舆情的价值变化，提高对舆情的走势进行预知管理。此外，相关领域的活动流程如果体现出了生命周期的流程，那么相关领域活动的生命周期与网络舆情的生命周期可以进行联系研究，从而揭示某一主题领域活动生命周期和其网络舆情生命周期的相关关系（胡昌平，1995）。

2. 基于施拉姆理论的舆情处置管理模型

研究信息交流的本质和规律的一种重要的方法就是研究信息交流模式（肖勇，2001）。比较著名的信息交流模式有申农-维弗的通信模型、拉斯韦尔的5W模型、施拉姆模型、维克利的S-C-R模型。申农-维弗通信模型为信息交流传播过程奠定了基础，拉斯韦

尔的 5W 模型关注交流的效果和信息本身，然而这两种模型的局限性在于忽略了“信息反馈”的重要环节，信息交流是单向的。施拉姆模型和维克利的 S-C-R 模型强调信息交流时交互的、双向的，更符合现实中人类信息交流情况（孙帅，2014）。

“施拉姆第三模型”是美国传播学家威尔伯·施拉姆所提出的大众信息传播与交流三种模式的第三种，1954 年，美国传播学家威尔伯·施拉姆对申农-维弗信息交流模式加以修正，将“反馈”引进到信息交流与信息传递中，将原先单一的信息交流模式变成双向互动的过程。按照施拉姆的观点，信息交流中的主体之间是相互影响的，各主体都必须将想要表达的意义制成代码，传递给对方，同时须将对方传送来的信息译码作出解释以产生意义，具体模型如图 4-3 所示。

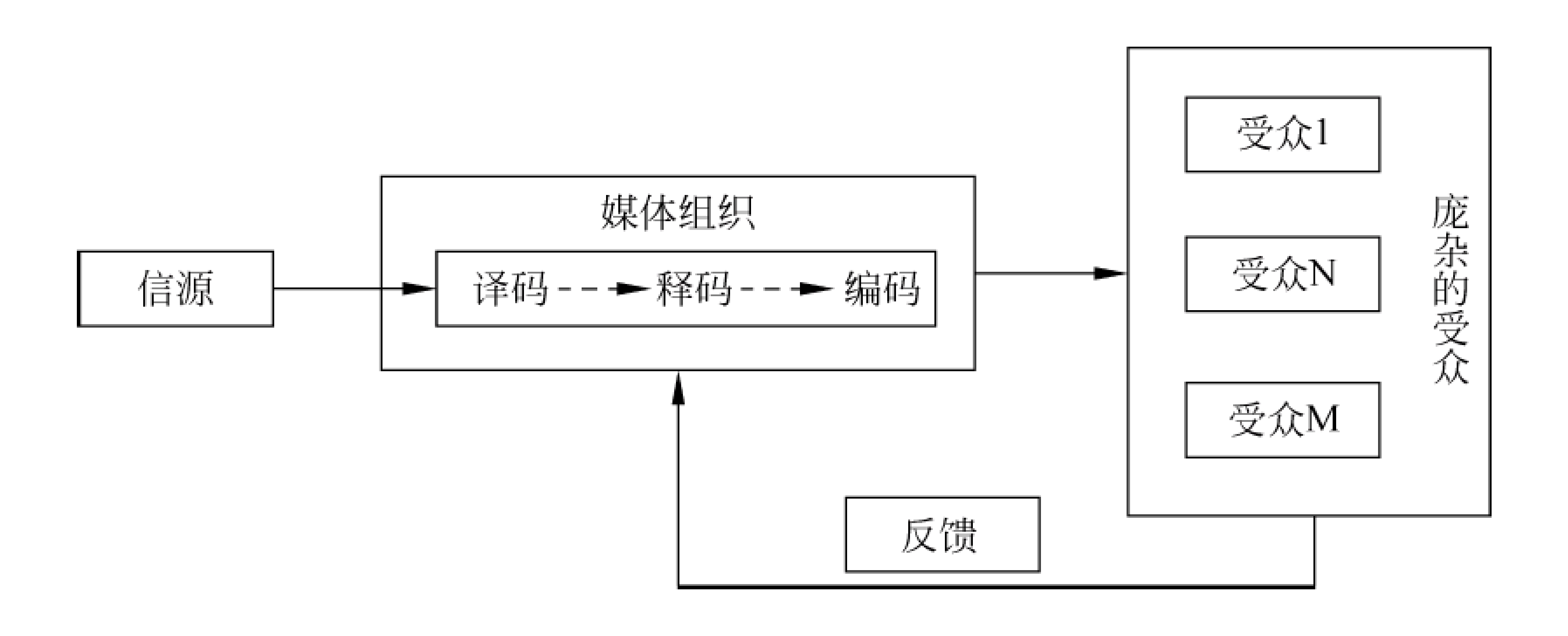


图 4-3　施拉姆信息交流第三模型

企业网络舆情中的信息流转结构符合信息交流的基本内涵。

首先对信息交流绝对变化与相对静止辩证平衡的体现（娄策群，2007）。企业网络舆情具有高度的动态复杂性。但从宏观上看，企业网络舆情在单一时间节点、空间节点的表现存在相似性。

当舆情事件发生后，企业网络舆情反应迅速，经过部分最初少量的受众将信息复制传递给更多更庞杂的受众，造成舆情事件消息的进一步扩散和关注度迅速升温，这是最典型的企业网络舆情的传播途径，而信息交流也同时在传播者与接受者之间发生。

其次对强调信息交流中信息主体与信息本身整体性的体现。企业网络舆情涉及多个参与群体，各参与主体与突发事件信息交换之间关系紧密，形成一个系统完整的整体。

最后是强调信息交流中正面价值与负面价值的辩证共存。企业网络舆情在其传播过程中，多方参与群体往往会围绕某个问题形成多方态度。但在舆情事件中，由于各主体之间信息的高度不对称，极易形成以“谣言”或“流言”为代表的各类消极舆情。因此，在舆情事件信息交流中，显著存在着积极舆情与消极舆情、正面价值与负面价值的共存现象。

总而言之，企业网络舆情信息交流的基本特性包括（惠志斌，2012）：传播过程复杂，涉及主体多元，信息传播以双向、多向的信息交流互动为主，政府在其中的作用被进一步削弱，网络舆情复杂程度空前提升等。信息交流的主体多元化及隐蔽化，若不能进行有效地引导和控制，对于舆情事件的有效处理将留下隐患。因此，加强对关键用户的梳理和引导，对于企业舆情管理部分做好突发事件的网络舆情管理工作至关重要。

3. 基于价值累加理论的舆情处置管理模型

针对群体性事件的发生机制，美国社会学家斯梅尔塞于 20 世

纪 60 年代提出了一个经典模型——“加值理论”(value-added theory)(曾润喜和徐晓林,2010),其对于解释群体性事件的发生机制仍然具有较为成熟的指导意义。该理论认为所有的群体性行为、社会运动甚至革命的发生,都是由“结构性诱因”、“结构性紧张”、“一般性信念”、“触发事件”、“社会动员组织”和“社会控制失效”六个因素的“加值”而发生的。所谓“加值”,指上述六个因素的孤立、并列或乱序出现并不足以导致群体性事件的发生,但当其按照一定的顺序出现时,它们的贡献就会被累加放大,从而大大增加群体性事件出现的可能性,如图 4-4 所示。

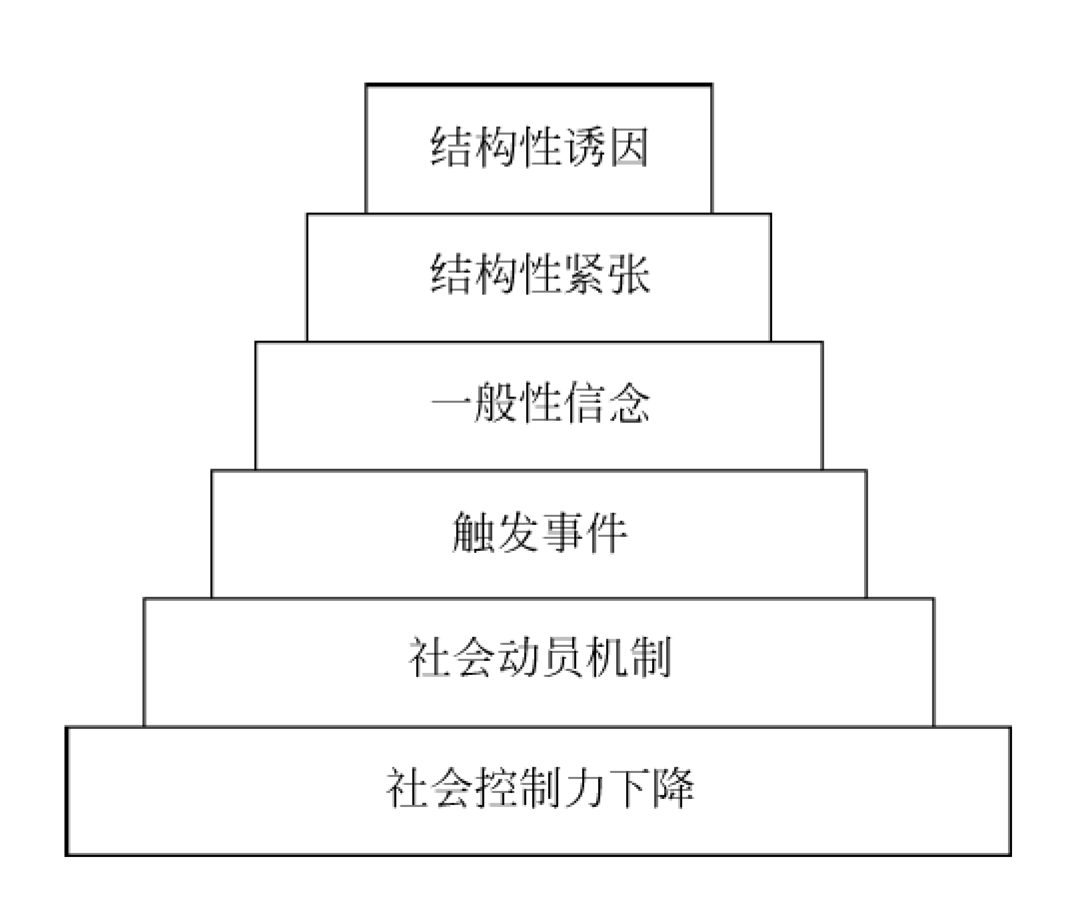


图 4-4　斯梅尔塞“加值理论”模型

对于解释企业网络舆情的发生机制,斯梅尔塞的“加值理论”同样具有借鉴意义。公共事务(产品、服务)是构成网络舆情的核心要素,企业网络舆情的本质是舆情事件所引发公共事务(产品、服务)及民众、企业、政府等利益攸关方围绕舆情事件形成的现实

矛盾在网络中的反映(林敏,2013)。

首先,企业网络舆情的社会投射性,完成了将群体性事件发生机制应用于企业网络舆情事件发生机制的逻辑跳跃。社会存在决定社会意识,作为社会意识的重要组成部分(吕嘉,2001),网络舆情的一切表现都植根于现实世界中的社会运动,不仅每一个网民都拥有现实存在的社会角色,网络中的每一个话题、事件,也都是社会矛盾的真实投射。

其次,企业网络舆情的系统规律性,保证了将孤立网络舆情事件发生机理适用于网络舆情整体生成规律的逻辑跳跃。从系统论的角度看,“企业网络舆情”是一个典型的信息生态系统(康伟,2012)。这一系统兼具内部复杂性和外部统一性的双重特性。

最后,企业网络舆情的动态开放性,实现了将企业网络舆情发生的阶段性规律适用于企业网络舆情演化全过程的逻辑跳跃。企业网络舆情中存在着显著的起伏与涨落现象,其起伏与涨落之间也存在着明显的因果联系,而企业网络舆情的开放性,允许身处不同地域的、持有不同观点的网民共同参与到某一事件之中。因此,利用“加值理论”对企业网络舆情微观事件发生规律的剖析就是对企业网络舆情全过程的深刻审视。

4.4 本章小结

在社会网络大数据的环境下，合理把握企业舆情，可以为企业提高经营管理绩效、引导舆情传播提供基础。本章首先研究了基于模糊推理法的舆情预警管理模型、基于情感倾向分析法的舆情预警管理模型、基于 Web 数据挖掘法的舆情预警管理模型以及基于网络舆情指标分析法的舆情预警管理模型四种分析预警管理模型，接着讨论了基于生命周期理论的舆情处置管理模型、基于施拉姆理论的舆情处置管理模型以及基于价值累加理论的舆情处置管理模型三种干预处置管理模型。

思考题

1. 简述基于情感倾向分析法的舆情预警管理模型内容。
2. 简述基于网络舆情指标分析法的舆情预警管理模型内容。
3. 举例说明基于施拉姆理论的舆情处置管理模型的应用。
4. 举例说明基于价值累加理论的舆情处置管理模型的应用。

第 5 章

数据平台和系统结构

本章学习目标

- 了解互联网数据获取的过程。
- 熟悉数据分析系统的主要功能模块。

5.1 数据获取

对互联网的新闻、论坛、博客等信息进行定时抓取，抓取的网页由 URL 来进行标识，URL 的处理的步骤为：

(1) 把一个初始的种子 URL 放入等待队列。

(2) 处理线程从等待队列中取出一个待处理 URL。

(3) 处理线程通过 Internet 抓取 URL 所代表的 Web 页，调用

页面处理模块处理页面。

（4）抽取出页面中的URL放入等待队列。

（5）把处理成功的URL放入处理成功队列，把处理失败的URL放入处理失败队列。

在完成一次抓取之后把处理的URL放入处理成功队列或者处理失败队列，然后重新到URL清单中取得一个URL，重复上面的过程。为提高获取信息的效率，在程序中的爬虫为多线程实现，多线程涉及爬虫任务分配的问题，可以用排队论中的结论来分析形成任务分配算法。在排队论，有典型的M/M/1问题和M/M/c问题，其中，第1个M代表网站名单到达过程是一个Poisson过程；第2个M代表爬虫引擎工作时间负指数分布，且各网站的收割时间相互独立（网站的收割时间由以下几个方面确定：爬虫引擎出去的带宽、路经的互联网各段的带宽、被收割网站进去的带宽、被收割页面的长度、同时访问该网站的客户数目、该网站服务器的速度等）；第3个位置的c代表有c台爬虫引擎。

此外，在爬虫引擎的工作中，我们假设系统的容量没有限制。由于网站名单是存储在一个文件中的，使用时导入内存，虽然文件长度和内存都有限制，但是，多几千、几万个网站名对计算机系统基本没什么影响。而几万个网站的情况也不常见，故我们这条假设在实践中基本可以满足，没有必要担心。还有，我们假设顾客源是无限的，即互联网上的网站名的数目是无限的。在实践中，这条假设也不难满足。

设网站到达规律服从 Poisson 分布，爬虫引擎对网站的收割时间服从负指数分布。设网站的平均等待时间为 W_c。可以证明：当 $c>1$ 时，$W_c>W_1$。也就是说，采用多收割线程，单个收割清单的办法比采用多爬虫线程，每个爬虫引擎各自有一个收割清单的办法的平均等待时间要少得多。所以，多线程爬虫引擎的任务分配方法，就是采用单收割清单，由一个爬虫引擎管理员顺次管理的方法。目前搜索引擎界普遍采用的这种方法，在理论上是具有其合理性的。

基于主题爬虫的设计，我们构建了一个主题爬虫示范平台，全称为金融信息爬虫服务系统，系统采用 Java 作为开发语言，该软件具有如下的一些特点：

(1) 可定期从互联网上自动获取金融相关信息，不需要人工的过多干预可完成。

(2) 可按照新闻标题、日期、股票代码和新闻类别对金融信息进行分类存储。

(3) 结合中文分词和语义理解可以很好地保证系统搜集金融信息的准确率。

社会网络大数据环境下企业舆情管理模式是一个较新的领域。我们运用社会网络计算、计算机智能、行为科学等相关理论和方法，来综合地分析研究社会网络大数据环境下企业舆情管理的优化问题。

5.2 数据平台

在进行企业舆情分析之前，我们需要完成社会网络数据的采集和存储。社会网络数据的采集就是编写抓取程序，从社会网络上采集用户原创信息或转发和评论的信息。编写抓取程序时，通常从一个或若干个网页的初始 URL 开始，获得初始网页上的 URL 列表，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入待抓取队列，直到满足系统的停止条件。经过初步去噪整理后，计算机服务器会将源网站名、网址及其发布时间存入本地数据库。

如图 5-1 所示给出了一个在使用计算机服务器对互联网信息舆情数据的收集和分析系统的示意图，相应的企业舆情管理平台，如图 5-2 所示。

我们针对社会媒体计算的跨学科特点，提出了社会媒体计算通用实验平台框架，并对平台的各个层次的关键核心问题进行研究。社会媒体计算实验平台是一个集成数据分析与处理、建模与仿真、预测与监控功能为一体的通用社会媒体计算实验平台，平台由对象层、数据层、模型层、分析层、应用层五部分组成。

(1) 对象层：对象层是社会媒体计算的研究对象。根据社会媒体计算的概念，从对象视角来看，社会媒体计算是要为社会进行计算。从根本上来讲，社会媒体计算要为现实社会服务，但从逻辑



图 5-1　数据库管理界面

上来讲，主要包括信息网络社会和复杂经济社会，以及从中抽取出来的人工社会。

（2）数据层：用户不但可以从互联网获取数据信息，还可以参与互联网活动，用户留下了大量社会“足迹”，对这些社会足迹进行收集与集成，为社会媒体计算提供基础的数据信息。重点围绕海量多源数据集成、多源数据融合，以及数据质量等海量数据信息处理的核心关键问题进行研究。

（3）模型层：对系统的微观建模是社会媒体计算的前提和基础。通过建模，对社会媒体计算系统中所涌现的宏观现象或情景

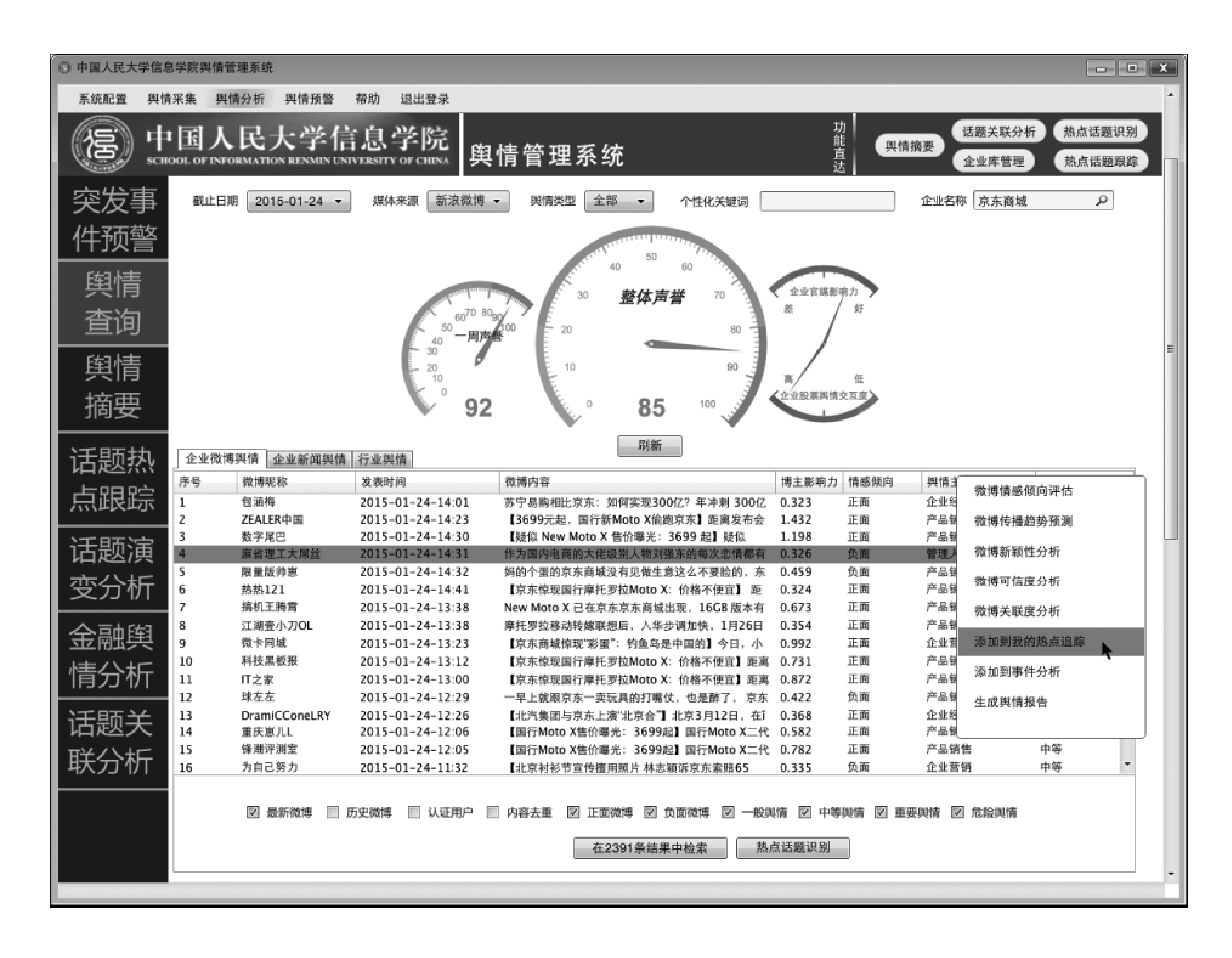


图 5-2　社会网络舆情管理平台

进行生动、形象地展示或可视化，对实验参数的可控化进行不同的情景再现，并按研究需求进行适时调整以方便决策者及时讨论形成有效的决策方案。

（4）分析层：社会系统行为是由个体相互交互而成，通过对个体行为分析研究系统内部结构是社会媒体计算的理论基础。分析层对系统结构进行微观分析，为社会媒体计算应用研究提供基础。

（5）应用层：作为社会媒体计算平台的最高层，应用层是把社会媒体计算的理论与方法应用到实际的社会、经济系统中，为政府、企业等相关部门决策提供依据。重点对社会网络传播、社会网

络计算、知识管理等领域进行研究。

5.3 系统结构

社会网络企业舆情建模和预警可以分为5个模块，如图5-3所示。下面对各模块分别简要说明。

1. 模块1：舆情大数据计算机基础模块

该模块主要包括基于Scripy的大规模实时数据获取平台的构建技术、基于Apache Spark的舆情大数据管理系统构建技术以及相关社会网络图像处理框架构建技术。该模块分别从舆情的获取、存储管理以及分析三个方面对社会网络大数据环境下的舆情管理方法所需的计算机技术进行了实现，是理论分析的现行技术基础。

2. 模块2：基于舆情传播内容的分析管理模块

该模块主要包括含短文本在内的舆情情感倾向分析技术及突发事件图像分析技术。基于内容的舆情分析基础是从舆情载体本身出发，对舆情所含有的静态信息进行分析，通过该管理技术，实现对于社会网络的静态监控与管理的目的。

3. 模块3：基于舆情传播结构的分析管理模块

该模块主要包括权威用户模型分析理论、串并联关键路径分析理论、用户群体行为模型。基于舆情传播结构的分析管理模块是从舆情的动态传播角度进行出发，对舆情传播的网络结构特征

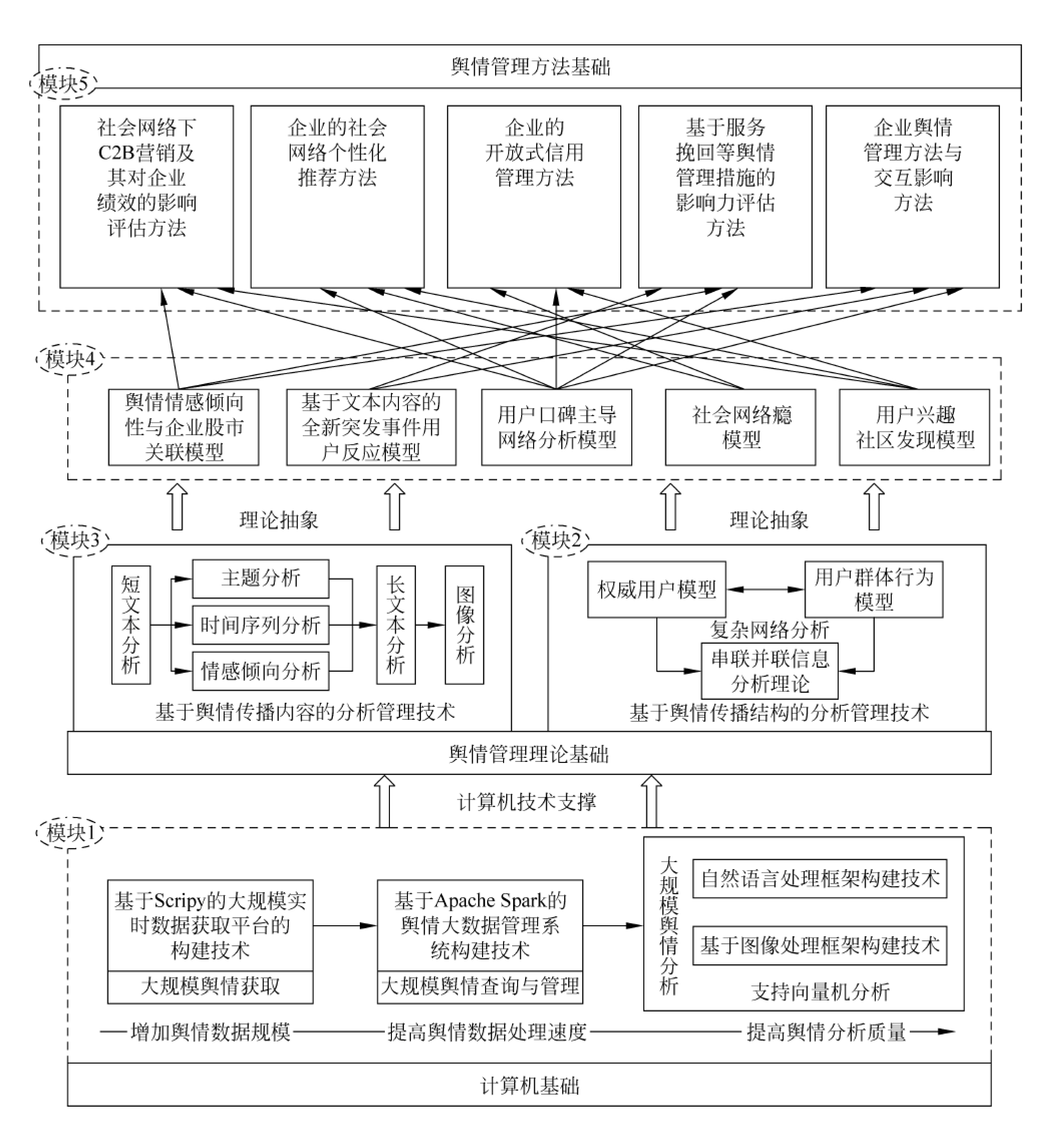


图 5-3　模块

进行分析，构建舆情传播的动态分析理论。基于结构的舆情分析理论模块是社会网络大数据环境下舆情分析的必要理论之一。

4. 模块 4：社会网络大数据环境下的舆情管理理论模块

该模块主要包含口碑媒体主导网络关键路径模型、舆情用户兴趣社区发现模型、社会网络瘾模型、舆情情感倾向性与企业股市关联模型、企业全新突发事件预警模型等。该模块是对模块 3 中模块的多角度总结，是在模块 3 的基础上派生出当前舆情管理方法所需要的基本理论，为社会网络大数据环境下的舆情管理方法的研究奠定基础。

5. 模块 5：社会网络大数据环境下的舆情管理方法模块

该模块是在模块 4 的基础上，研究用户群体行为对企业业绩影响的 5 个企业舆情问题，包括 C2B 营销方法、企业的社会网络个性化推荐方法、企业的开放式信用管理方法、企业舆情管理方法与交互影响方法、基于服务挽回等舆情管理措施等优化管理方法。通过模块 4 中的基础理论的不同组合，来灵活处理模块 5 中的不同问题，使得研究系统化。换句话说，通过对模块 4 中社会网络大数据环境下舆情管理基本理论的研究，并进行丰富的组合，从而来解决社会网络大数据环境下企业舆情管理所遇到的大多数问题。

5.4 本章小结

要进行企业的舆情分析，首先要完成社会网络数据的采集和存储。社会网络数据的采集就是从社会网络上采集用户原创信息或转发和评论的信息，经过初步去噪整理后，计算机服务器会将源

网站名、网址及其发布时间存入本地数据库。社会网络企业舆情分析系统可以分为五个模块，分别是舆情大数据计算机基础模块、基于舆情传播内容的分析管理模块、基于舆情传播结构的分析管理模块、社会网络大数据环境下的舆情管理理论模块及社会网络大数据环境下的舆情管理方法模块。

思考题

1. 简述互联网数据获取的过程。

2. 举例说明社会网络大数据环境下的舆情管理方法模块的主要内容。

第 6 章

企业网络舆情管理的计算机技术

本章学习目标

- 熟悉基于文本内容的企业网络舆情管理的技术。
- 熟悉基于图像内容的企业网络舆情管理的技术。

6.1 基于文本内容的企业网络舆情管理的技术

1. 文本情感分析算法

随着网络以及新社交应用的快速发展，大量网络用户每天都会发布并传播高达上亿的信息。这些海量的文本信息中，有很大一部分是表达用户观点倾向和情感信息，这些情感文本信息是非常宝贵的意见资源，包含着人们对社会各种现象的不同观点和立

场。情感分类将评论自动分类为正面或负面的意见，目的是挖掘消费者对某种产品或服务的评论。不仅强烈地影响消费者的决策过程，对于管理者考虑品牌建设，产品开发和质量保证也有重要的意义。

目前情感分类的实证研究不仅突破了行业限制、产品限制、地域限制也突破了语言的限制，其被运用在多处，例如旅游业、餐饮业、影视的评价等等，不仅仅局限在英文的评论中，有学者也提出了中文的情感分类算法，叶强等(2007)也对粤语的情感分类进行了实证研究。

利用计算机智能技术，包括自然语言处理、机器学习、文本挖掘等(梁循，2006；He 和 Zhou，2011；Neviarouskaya 等，2011)，我们可以将各种人类情感、意见或观点由社会网络的机构化文本信息转化成定量的数值数据信息。情感倾向性分析比较系统的研究工作开始于基于监督学习方法对电影评论文本进行情感倾向性分类和基于无监督学习对文本情感倾向性分类的研究(王超等，2009)。学者们分别使用朴素贝叶斯、最大熵和支持向量机(Liang 等，2008；Liang，2010；Liang 和 Ni，2011)技术将文本情感倾向性分为负向和正向两类。近几年来，文本情感倾向性分析(也称文本情感计算、文本褒贬性分析)已经成为国内外计算机科学界的一个热点研究方向。目前，情感倾向性分析已经被运用于英文、中文等许多种文字，并获得了很大程度的发展，特别是在在线评论的情感倾向性分析上获得了很大的进步，基于在线评论文本的情感倾向

性分析的准确率最高能达到90%以上(杨源等,2012)。

为了找出评论的情感倾向性,我们需要借助的智能手段包括自然语言处理、机器学习、文本挖掘等(梁循,2006;He和Zhou,2011;Neviarouskaya等,2011)。

显然,当舆情的情感倾向性值为正时,表示对企业有利,反之为不利。因此,情感倾向性值可以认为是一个企业舆情的基于社会网络的基本测度,也就是说通过文本挖掘与自然语言处理技术,我们可以得到企业在社会网络大数据环境下情感倾向的定量式结果。在本书中,我们定义情感倾向的取值范围为[0,1],贬义议论的取值范围为[−1,0],于是,情感倾向的取值范围为[−1,1]。相应地,企业社会网络舆情测度值的取值范围也为[−1,1]。

在本研究中,我们可以挖掘到的评论信息按照标题存储为不同的信息文档。为了使用情感倾向性,需要计算单个信息文档的情感倾向性值。我们利用中科院的分词工具ICTCLAS对每个信息文档进行分词处理,将汉语句子分解成一个个有意义的词汇$\{w1,w2,w3,\cdots,wn\}$,其中n为词汇数量。借助Hownet词典,我们得到了正面词典、负面词典、否定词典以及另外6个修改词典的词汇褒贬强度值vi,其中正面词典中有4566个词汇,负面词典有4370个词汇。除了正面词典和负面词典外,其他7个词典都是用来修饰正面和负面词汇用的。例如,如果否定词出现的正面或负面词前的话,则对其取反,$vi=(-1)*vi$。而修饰词典根据他们的修饰程度级别分别赋予一定的强度。接下去,判断wi前面的两个

词汇 precede1 和 precede2 和后面的两个词汇 post1 和 post2 是否在修饰词典中,如果出现,则将 vi 乘上修饰词的权重。整个信息文档的情感倾向性值就是所有词语的 vi 之和,从而得到用户评论的情感倾向性 W。

情感分类算法可以分成几大类。朴素贝叶斯算法,支持向量机算法和基于字符的 n-gram 模型这三种标准的监督机器学习方法都被证明能够进行很好的情感分类。但在不同算法的精确度上,在不同的实证环境下略有区别。

Ye 等(2009)在研究世界七个主要旅游目的地的网上评论时,采用朴素贝叶斯算法,支持向量机算法和基于字符的 n-gram 模型三个监督的机器学习算法发现,训练有素的机器学习算法可以进行很好的分类。虽然最近的一些研究已经开始进行旅游博客的内容分析,但是复杂的网络挖掘技术仍需要融入旅游博客的分析中。这项研究填补了旅游评论中的空白,进行了基于旅行目的地的在线评论的情感态度的自动分类。

Zhang 等(2011)在研究互联网上餐厅的粤语评论的情感分类时,将标准的机器学习技术朴素贝叶斯和支持向量机运用在网上粤语写的餐厅评论当中,以自动分类用户的评论是正面的还是负面的。也对分类性能的影响进行了讨论。其发现,分类的精度取决于分类模型和特征选择之间的相互作用。朴素贝叶斯分类器的精度与支持向量机相当甚至更好。双字词组比单字和三字词组能够更好地捕捉粤语情感倾向。

除了将标准的监督机器学习算法运用在实证的研究当中，国内的学者还根据英文的语义指向算法进行了改进，提出了适合中文情感分类的算法。

Ye 等(2006)在对电影的中文评论情感分类研究中，就着重于研究中文评论的分类算法。鉴于中文与英文的语言差异，英文情感分类的语义指向算法不能直接适用于中文的情感分类中，因此基于对中国电影评论的情感分类研究来改进语义指向算法以此探索中文的情感分类方法。根据这个结果发现，中文评论所改进的语义指向算法的性能与先前的英文评论分类研究相比，是可以接受的。数据集的测试还表明，参考字对的选择和语义指向的临界值对评论分类的语义指向有着重要的影响。分词方法被引入情感分类过程中，是中文和英语在情感分类中的主要区别之一。这项研究还介绍了最佳参考字对的选择和最佳语义指向的临界值的选择。研究过程表明，当语义指向方法应用于不同领域、不同产品的情感分类，参考字对和临界值的选择是不同的。

在进行实证研究时，无疑会受到领域以及产品或者服务类别的限制。例如在研究世界七个主要旅游目的地时，就应当扩大目标数，虽然这项研究分析了西方国家热门的目的地，但在其他目的地所提出的分类方法的适用性仍然不明，因此值得今后的深入研究。由于消费者会经常改变他们对于目的地的看法和感知，因此还可以比较不同时期的分类结果做更为深入的纵向研究。

在对中文影评的改进算法的研究中，在提取短语时，只是简单

地借鉴了来自英语评论当中的两词词组的模式。但英语和中文的语言结构有许多的不同，如何能在中文评论挖掘的过程中，找到更为精准的文字，将是进一步研究的突破口，也将大大地提高情感分类的准确度。

此外，应该以中文为导向的情感分类算法做进一步的优化和推广，兼顾到不同地域和不同方言的适用性，尤其对中文的分词方法以及语言表述的准确性上带来更大的突破。

网络评论中也存在许多垃圾信息。这些信息的存在无疑会干扰基于情感分类的相关研究，应当注意对于垃圾信息的过滤。网络评论的情感也较为丰富，可以进一步延伸对于情绪强弱的判别，不应只局限于正、负两个方面。

在社会网络中，文本信息有其特殊性，就是都不太长，有的甚至只有几个字。所以，有些学者提出了短文本的问题，并提出了一系列处理方法（梁循，2012；詹志坚，2014）。

2. 短文本自然语言处理技术

短文本是相对于长文本而言的，它是指那些文本长度小于160个字符的文本，一般以微博、手机短信、网页评论以及聊天等形式存在。近年来，随着移动物联网及各类新兴社交网络的飞速发展，短文本成为人们获取信息的重要来源。短文本的研究已成为了目前的一个研究热点。短文本自然语言舆情分析是社会网络舆情分析的核心科学问题，其研究内容涉及短文本建模、相似度计算、话题挖掘、情感分析等。

1）短文本建模

短文本建模是对短文本舆情分析计算的基础。目前，对短文本的表示方法主要沿用原有传统的文本建模方法，主要包括：字符表示法和词袋表示法等。字符表示法是把短文本视为连续的字符并以字符作为基本的处理单元。词袋表示法把短文本看作一组特征词的集合，并通过特征词及权重表示为多维空间中的一个特征向量来表示短文本。词袋表示法是最常用的短文本表示方法，在传统的向量空间模型、概率模型中都大量采用了词袋表示法来对短文本进行建模。

2）短文本相似度计算

短文本相似度计算是短文本舆情分析的关键技术。通过对短文本的相似度计算并进行分类整理，进而可实现短文本主题提取。短文本相似度是一个重要且应用广泛的概念，但目前尚未对短文本的相似度有令人信服的准确定义。现阶段，使用较为广泛的短文本相似度定义为

$$\mathrm{sim}(A,B)=\frac{\log \mathrm{P}(\mathrm{common}(A,B))}{\log \mathrm{P}(\mathrm{distinction}(A,B))}$$

其中 common(A,B)和 distinction(A,B)分别表示短文本 A 和 B 的共同部分和不同部分（Lin，1998）。

基于语义词典的短文本相似度计算方法（Meng 等，2013）是目前运用最广泛且有效的方法。WordNet 是目前使用最为广泛的语义词典。基于大规模语料库统计的短文本相似度计算方法也是近

年来研究较多的方法，其主流有 LSA(latent semantic analysis)和 HAL(hyperspace analogues to language)。LSA 将词和短文本映射到语义空间，通过减少向量空间维度，减少计算量，提高检索精度。HAL 通过在大规模语料库中寻找词共现信息获取次或文档的语义信息。在短文本相似度计算中，LSA 较 HAL 有更好的效果(Schutze，1998)。基于描述特征的短文本相似度计算方法预先维护一个特征库，将预处理后的短文本映射到特征库中，得到短文本对应的特征向量，从而将短文本的相似度计算转换为特征向量间的相似度计算。基于互联网资源的方法(Sahami 等，2006)通过互联网丰富的资源，对信息量少内容简短的短文本进行扩充，使待计算相似度的短文本之间包含相同词项或相似的可扩展词语。岳云飞等(2014)依据微博短文本之间的关联关系扩充微博短文本特征信息，并将扩充后的特征词集采用 HowNet2000 映射到概念集，最后采用 Jaccard 相似度系数计算短文本间的相似度。

词语相似度计算是短文本相似度计算的基础。结合企业舆情分析，学者们提出了下面的中文词语相似度计算方法，

$$\mathrm{BaikeCatSim}(A,B)=\sum_{k=1}^{m}\frac{\mathrm{IIC}_k(\mathrm{LSO}_{A,B})}{\mathrm{depth}(A)+\mathrm{depth}(B)}$$

其中，m 表示词条 A 和 B 之间的最短路径条数，$\mathrm{IIC}(W)$表示节点 W 的信息内容，$\mathrm{IIC}(\mathrm{LSO}_{A,B})$表示词条 A 和 B 最近公共自节点的信息内容。$\mathrm{IIC}(W)$的计算公式如下，

$$\mathrm{IIC}(W)=1-\frac{\log|\mathrm{hypo}(W)+1|}{\log|\mathrm{ALL}|}$$

其中 hypo(*W*)表示节点 *W* 下所有节点的数量,ALL 表示百度百科所有节点数量。

3) 短文本话题挖掘

短文本话题挖掘是指通过对微博、在线评论等短文本进行基于内容(语义)的聚类分析,挖掘各种热点话题及其观点和立场。通过及时掌握人们对各种热点话题的观点和立场及对话题的发展进行分析,对国家、社会都具有重要意义,引起了学术界的高度重视,它是短文本舆情分析的重要目标之一。

目前,针对短文本话题挖掘方法主要有:

(1) 文本聚类法。采用文本聚类的方式进行话题挖掘。如路荣等(2010)利用 K-means 和层次聚类两层混合聚类算法,利用隐主体模型挖掘的热点新闻话题。

(2) 传统模型的改进。针对短文本的特征,改进传统话题模型潜狄利克雷分配(latent Dirichlet Allocation,LDA)模型,利用所建立的新的话题模型来抽取短文本话题。如 Ramage 等(2010)建立了一种半监督学习模型 L-LDA 满足用户在 Twitter 上个性化的信息需求。

(3) 通过分析短文本的内容,自动产生短文本内容的摘要提取。Inouye 等(2010)提出一种用多条句子来描述微博热门话题的方法,克服了一条句子对话题描述不完整的缺陷。Sharifi(2010)实现了对微博话题的描述,使用户可以实时并且准确地得到微博中的热门话题。

4）短文本情感分析

短文本情感分析主要是指从短文本信息中识别主观信息，挖掘用户对产品、新闻、热点事件等评论信息所持有的观点和态度，这也是舆情分析的重要目标之一。短文本尽管信息含量少，但其情感信息丰富。短评论所缺失的信息一般是产品的主题和特征信息。短文本内容越小，其情感分析效果越好。这是因为较长的文本其所包含的不利于情感分析得噪音越多。早期对于短文本的研究大多集中在主题挖掘上，而对于情感极性分析较少。微博等网络中的文本不仅短小，且数量极其众多，同时包含丰富的极性词，极其利于情感分析。

国外对短文本的研究相对较早，Read(2005)详细论证了表情符号对情感分类的作用。Go等(2009)采用无监督指导的朴素贝叶斯、最大熵和支持向量机三种机器学习方法，使情感倾向判别的准确率超过了80%。Pak等(2010)组织标注了Twitter微博文本情感极性数据集，实现了基于朴素贝叶斯、支持向量机和条件随机场的情感分类器。由于微博是近几年才在国内发展起来的新型社交媒体平台，所以国内方面针对微博等短文本的情感分析研究还相对较少。谢丽星等(2012)对基于表情符号的规则方法、基于情感词典的规则方法和基于SVM的层次结构的多策略方法进行了深入的研究，结果表明基于SVM的层次结构多策略方法效果最好。

综上，短文本由于特征向量的维度过少，在整体特征矩阵中不可避免地出现极度稀疏的问题，即每个短文本样本中，只有极少数

的维数上是有取值的。由此，给短文本的处理带来了极大的不确定性和困难。解决这种稀疏性主要从两个方面入手：一是通过降低整体的特征向量维度来避免稀疏性问题。信息增益方法、CHI方法以及互信息方法，其他的方法有潜在语义索引（Chen 等，2003）方法、基于聚类重心数据降维（Kim 等，2005）的方法等。这些降维的方法或者需要计算大矩阵的特征值和特征向量，或者需要对数据进行频繁的聚类迭代分析，其计算复杂度和计算时间都比较大。二是通过各种方法扩展短文本的信息，从而提高短文本自身的向量维度。短文本存在着数据稀疏性及上下文缺失的情况，需要用某种方法来补充和扩展信息。如 Wang 等（2007）利用 WR-kmeans 聚类方法综合相关手机短消息解决相似短文本发现问题；Fan 等（2010）利用特征扩展和控制模型，有效提高短文本的分类精度；Adams 等（2008）利用 WordNet 解决即时聊天信息话题检测与抽取问题。

进一步，有效地监控这些在线文本可能也是非常重要的。有时候在线文本中负面的“流言”可能对企业造成后果严重的问题，及时发现并加入正面的帖子，就可以有效地防止问题的发生。近年来，舆情监控成为国家管理互联网的一种必要技术（Liang 等，2012）。所以，上述应用相当于“企业级”的舆情分析监控。

3. 应用：基于微博的文本内容的突发事件及全新突发事件预警方法——用户在紧急状态下的行为分析

对常见突发事件，人们已经有了比较成熟的预警技术。对全

新突发事件的技术，也有了一些进展。例如，在梁循和申华(2012)专利中，互联网上的新词可以进行甄别，并判断是否是突发事件。

假设我们对企业A的舆情的全新突发事件进行监控。

首先，我们利用微博开放平台，抓取微博信息。在此基础上，实现对用户所关注对象微博发布信息的提取，并将提取的语片级微博进行分词。其次，根据微博长度按照分类规则循环累计式的找出每一语片级微博中的关键词，将其放入"候选关键词"队列，将得到的候选关键词分别与《企业A突发事件字典》比较，若无任何一个关键词在《企业A突发事件字典》中，则放弃本微博；若微博关键词存在于《企业A突发事件字典》中，则将本微博中不包含于《企业A突发事件字典》的候选关键词列入"全新突发事件候选关键词"队列。依次循环的将设定时间内的所有微博进行分析提取。最后按照关键词出现的频率排序，并对前十个"全新突发事件关键词"进行关键词共现分析，最终可视化展示出所得结果。

具体步骤是：

步骤1，首先利用微博开放平台和API接口技术。根据所需，指定将要抓取的用户类别和微博发布时间段，利用Java语言编写的抓取程序，实现指定微博的抓取。并存储进入数据库。

步骤2，对该条微博进行分词。

步骤3，根据微博总字数长度不同选择提取关键词。

步骤4，将本条微博"候选关键词"队列中的词语与《企业A突发事件字典》进行比较。

结果 4.1：如果“候选关键词”队列中，无任何一个词语在《企业 A 突发事件字典》中，则放弃本条微博，进行下一条分析。

结果 4.2：如果“候选关键词”队列中，出现了《企业 A 突发事件字典》中的词语，则将不在《企业 A 突发事件字典》中的其余关键词列入“全新突发事件候选关键词”队列中，并以向量组形式存储在数据库中。

步骤 5，处理完指定的所有用户在预设时间内的所有微博后，提取出“全新突发事件候选关键词”向量组。并对每一个向量中第一个元素，即“候选关键词”出现的频率进行统计，按照统计结果由大到小的词频顺序进行排列，提取出词语出现频率最高的前 10 类“全新突发事件关键词”。至此得到所需的全新突发事件的关键词。

步骤 6，得到的 10 类“全新突发事件关键词”与向量组中第二个元素及以后的所有词语进行关键词共现分析。并通过共现分析得到全新关键词语突发事件之间的内在联系。

4. 社会网络推荐技术

社会网络中推荐的基本思想是，具有相似兴趣爱好用户会对同一商品表现出相似的偏好。根据此思想，要对用户进行推荐，先要确定其邻居（相似）用户，然后再综述其邻居的偏好。所以，对某一用户的推荐，实际是取决于其他用户的偏好，用户与用户之间形成一种自助式、协同式的社会推荐模式。

传统推荐方法大多是基于内容的，首先需要对用户和候选推

荐对象分别进行建模表示，然后把用户与产品进行匹配。也就是说，对某一用户的推荐，不需要参考使用其他用户的偏好信息。

研究基于社会网络中推荐的价值创造与运作创新（Tsai 和 Ghoshal，1998），就是运用社会网络中推荐的方法，通过研究用户的社会网络中活动的行为模式，利用计算机在社会网络上获取客户对推荐的反应，观察企业在生产、供应满足目标客户需要的产品或服务的一系列业务活动及其成本结构的新变化，丰富企业舆情的管理理论与方法。

Web 已成为人们获取信息的一个重要途径，然而“信息过载”使人们在日益多样的信息类型中迷失，推荐系统可帮助用户有效地解决网络上的“信息迷失”问题（曾春，2002；刘建国，2009）。个性化推荐是根据用户不同的需要、习惯、兴趣、动机、信念等个性化因素，因人而异地向用户提供差异化的产品或服务来满足用户的个性化需求（刘建国，2009；张秀伟，2013）。其核心就是建立用户与信息产品之间的关系，通过收集和分析用户信息来研究用户的潜在偏好特点和行为模式，建立个性化的用户模型，再根据用户的个性化模型将用户所需的信息传送给用户，实现个性化信息推荐，它是目前解决信息过载问题最有效的工具。个性化信息推荐系统中的核心部分是智能推荐技术，因此，对于推荐技术的研究一直是研究者们关注的热点和重点。目前为止，主要的推荐技术包含基于内容过滤的推荐技术；基于协同过滤的推荐技术；基于人口统计信息的推荐技术；基于经济的推荐技术；基于知识或情境的推

荐技术；基于关联规则的推荐技术；基于社会化标签的推荐技术；基于社会信任的推荐技术等。下面重点选取与企业网络舆情密切相关的五种推荐技术进行介绍。

1）基于内容过滤的推荐技术

基于内容过滤的推荐技术是信息过滤领域所派生和延续的一个分支，主要通过一些学习算法获取和更新用户偏好，如决策树、神经网络等学习算法；基于内容过滤的推荐技术要求对每个 i 及每个用户 u 进行描述，这些描述是基于一个特征空间而给定的，即每个 i 和一组特征（或属性）$(f_{i1}, f_{i1}, \cdots, f_{in})$相联系，而每个 f_{ij} 属于第 j 个特征的可取值的集合 F_j，相类似的，用户也被一组在 $G_1 \times G_2 \times \cdots \times G_m$ 集合上表示的特征所描述。根据这些描述，基于内容过滤的推荐方法通过计算用户偏好的表示与信息资源之间的相似性来得到用户对某个信息资源的偏好值。因此，基于内容过滤的推荐技术的关键问题是相似度计算，对于信息资源采用矢量空间模型表示的方法来说，通常采用的相似度计算方法是余弦度量法。基于内容过滤的推荐系统其优点是简单、有效，缺点是难以区分资源内容的品质和风格。它一般只能也就是信息内容同质的信息，不能为用户发现新的感兴趣的资源，所以基于内容过滤的推荐技术无法为用户提供新颖的推荐。

2）基于协同过滤的推荐技术

当前，协同过滤推荐技术是最流行、最广泛实现和最成熟的推荐技术。协同过滤推荐是根据用户的相似性来推荐资源，在协同

过滤推荐系统中,典型的用户偏好是由信息和它们被评分值所形成的向量组成,这个向量将随着用户与系统的不断交互而逐渐增大。在某些情况下,评分的值可以是二元的(喜欢和不喜欢)或者是用一定范围的实数值来表示用户的偏好程度(蔡登,2002)。随着 Web 2.0 时代的不断推进,互联网信息越来越丰富,尤其是用户的在线网络行为成为研究者们关注的内容(金淳,2013;王伟,2014),王伟(2014)通过产品的静态属性和用户的在线评论行为构建用户的偏好信息,以此来提高推荐的准确度和召回率。协同过滤推荐技术最大的优势是在于它完全可以脱离任何被推荐对象的机器可读的表示形式,甚至在一些复杂对象上依然工作得很好,协同过滤推荐技术一般采用机器学习理论中的最近邻技术,利用用户的历史喜好信息计算用户之间的距离,然后利用目标用户的最近邻对资源信息评价的加权平均值来预测目标用户对特定资源信息的喜好程度,系统会根据这一喜好程度来对目标用户进行推荐。

3)基于关联规则的推荐技术

基于关联规则的推荐技术是通过设定的规则为用户实现推荐。规则可以由用户定制,也可以利用基于关联规则的挖掘技术来发现;利用规则来推荐信息依赖于规则的质量和数量,基于规则的技术其缺点是随着规则的数量增多,系统将变得越来越难以管理;同时规则的形成过程需要的时间很长。一个规则实际上就是一个 if-then 语句,规则可以利用用户个人静态属性信息来建立,也可以利用用户动态信息来建立。实现信息推荐的工作过程是:

首先根据当前用户已阅读过的感兴趣的内容，通过规则推导用户还没有阅读过的感兴趣的内容，然后根据规则的支持度（或重要程度），对这些内容排序后展现给用户。

最常见的基于规则的推荐技术的应用领域是电子商务推荐系统，它采用数据挖掘技术中的关联规则方法对用户的购买记录进行分析，通过已购商品用户的商品购买的频繁集合推导出一些商品的购买规则，系统应用这些规则为其他用户实现推荐任务。当前，此种方法已被应用到某些商业网站中。

4）基于社会化标签的推荐技术

随着社会标签系统的出现，网络用户利用标签可标注博客、文章、音乐作品、电影、图像、产品等资源，由于标签的低门槛使用和低认知水平要求的特点，吸引了不同层次用户的使用，因此自发地产生了很多描述信息。标签的词汇是由用户自由选择，从某种意义上将聚合出资源的特征。标签是用户根据个体需要直接标注于资源上并与他人共享，其体现了用户、资源和标签之间的多维关系。因此，社会标签提供了一个观察用户之间关系的视角，为反映用户偏好提供了依据，是实现个性化服务的重要途径（张海燕，2012）。

基于社会化标签的推荐技术根据用户在资源上的标注行为特征和标签规律确定相似的资源群或用户群，即利用标签来确定资源之间的相似性和用户之间的相似性，然后利用这些相似资源群或用户群所使用的标签进行推荐。基于标签的个性化推荐系统常

见的是利用标签形成加权标签向量为用户兴趣建模，其围绕着用户在资源上标注标签的频率、次数、出现的特点等信息，采用概率计算模型等方法统计这些信息的规律。一般把规律转换成可计算的量，用这些量来反映用户的偏好，然后完成个性化推荐系统中的协同过滤推荐算法。Firan（2007）提出了基于标签的音乐推荐系统，采用标签的频率表达用户偏好，标签不仅反映了音乐本身的类别和内容特点，而且还反映了用户对音乐的喜好。此系统应用在last.fm音乐搜索网站上，不仅表现出了基于曲目的推荐算法同样的推荐准确性，还具有多样性和新颖性。Denis基于标签的BM25相似性测量方法改进了协同过滤推荐算法的相似性测量方法，由此来增加推荐过程的准确性（Parra-Santander，2010）。

5）基于社会信任的推荐技术

随着社会网络的迅猛发展，社会网络的大数据时代已然深入人心。在线社会网络中大量用户之间的交互信息成为Web数据挖掘的重要来源，也给个性化推荐系统带来了福音。基于社会网络实现推荐的模型得到了研究者们广泛关注，研究者把社会网络中用户之间的联系视作用户之间关系的另一维度，这就构成了一类上下文感知的社会推荐技术。在线社会网络是现实物理世界的折射，因此，如果在线社会网络中两个用户之间的交互很频繁，这意味着这两个用户之间的关系强度很大，更进一步可以理解为用户之间的信任程度与交互关系是相关的（邹本友，2014）。换句话说，当两个用户之间的交互关系为正关系时，用户间的信任程度会

提交；反之，信任程度就会下降（Sherchan，2013）。在 eBay 和 Amazon 等在线电子购物网站中，用户间的信任是根据他们之间历史交易的反馈来获得，采用用户之间的信任度可以提高物品推荐的满意度（Ruohomaa，2005）。Xiang（2010）提出了一种无监督学习的方法来确定社会网络中信任关系的强度大小，用户间的连接强度体现社会网络中的信任关系，强连接关系表示用户间的比较高的信任关系，弱连接表示用户间的信任关系低。Zarghami（2009）在文中引入 T-index 的概念来估计用户之间的信任程度，并且根据用户之间的信任程度来给用户推荐新朋友。研究者们把朋友、用户之间的信任度引入到社会化推荐中，以此提高推荐的准确度和个性化的需求。

6.2 基于图像内容的企业网络舆情管理的技术

除了基于文本内容外，图像一样可以反映舆情。我们的实证研究表明，有图像的微博的转发率，要明显高于只有文字的。此外，有很多转发的微博，只是简单地把图像进行了转发。所以，要研究企业的舆情分析，图像舆情是不可缺少的。

在识别图像内容的研究中，网络上的图像由于它的特殊性，一直是一个研究的热点问题。Smeulders 等（2000）指出其根本困难在于计算机能从 0 像素中提取的底层视觉特征和用户在特定情境下对图像内容的高层语义解读之间存在语义鸿沟。

为了消除语义鸿沟，现有方法主要是利用一组已有的人工标注的图像训练数据和识别标签之间建立某种映射关系，再根据这种映射关系自动为待标注的新图像添加相关标签。目前公认比较有效的解决方案是量化尺度不变特征变换（scale invariant feature transform，SIFT）描述子得到的图像特征配合 SVM 为主的分类器来完成。

SIFT 是用于图像处理领域的一种局部特征描述子，该算法对图像的尺度缩放、平移、旋转变换、甚至亮度变化以及仿射变换都具有相当的稳健性。

下面，我们将 SIFT 算法应用于图像物体识别和特征提取领域，重点识别提取出图像中较为典型的企业 logo、企业领导照片、企业名称等特征，此外，我们可以收集大量已标记为企业的图像，针对每一幅图像，完成以下四个步骤。

1. 尺度空间极值检测

针对离线收集到的企业图像进行亚采样，将平滑和亚采样重复进行，就可以得到构成金字塔的一系列图像。如下定义二维高斯滤波函数，

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{\frac{-(x^2+y^2)}{2\sigma^2}}\quad（其中，\sigma 表示高斯函数的方差）$$

输入的 $N\times N$ 企业图像 $I(x,y)$ 在不同尺度空间下的表示，可以由图像与高斯核卷积得到 Gaussian 图像，

$$L(x,y,\sigma)=G(x,y,\sigma)\times I(x,y)$$

其中，称为尺度空间因子，其值越小表示图像被平滑得越少。大尺度对应图像的概貌，小尺度对应图像的细节。DoG 算子定义为，

$$D(x,y,\sigma)=[G(x,y,k^{\sigma})-G(x,y,\sigma)]\times I(x,y)$$

为了检测 $D(x,y,\sigma)$ 的局部极值点，需要将 DoG 尺度空间每个点与其相邻尺度和相邻位置的 26 个点逐个进行比较。若像素 (x,y) 是一个可能的 SIFT 关键点，则它必须在周围 26 个近邻像素（上一个尺度的 9 个点+同尺度的 8 个点+下一尺度的 9 个点）中是极值点。所有这样的局部极值点，就构成了一个 SIFT 候选关键点的集合。

2. 关键点定位

极值检测得到的企业图像的所有关键点，还必须通过两步检验才能确定关键点：一是它必须与周围的像素有明显的差异，即需要提出对比度低的关键点；二是需要剔除不稳定的边缘响应点（因 DoG 算子会产生较强的边缘响应）。

3. 关键点大小和方向匹配

为了使算子具备旋转不变性，采用梯度直方图来确定关键点的主方向。点 (x,y) 处梯度的模值和方向的计算公式为

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2}$$

$$\theta(x,y)=\arctan\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}$$

对于企业图像的每一个关键点，考虑它的邻近的一个邻域窗

口内点的梯度方向，直方图的峰值就代表了该关键点出邻域梯度的主方向，即作为该关键点的方向。为每个关键点指定方向参数，使得算子具备旋转不变性。

4. 生成 SIFT 描述符

为确保旋转不变性，首先将坐标轴旋转为关键点的方向。以一个关键点为中心，取 8×8 的窗口，将该窗口切成 2×2 的子窗口，统计每个子窗口中的方向直方图。

每一个子窗口的方向由其上 4×4 的小块的方向用之前的方法来决定。企业图像中的每个关键点方向由 2×2 共 4 个种子点的方向决定，一个种子点有 8 个方向的信息，则每个关键点就有 4×8＝32 维。

在实际计算过程中，为了增强匹配的稳健性，通常采用 4×4 共 16 个种子点来描述，这样企业图像中每个关键点就有 16×8＝128 维的数据，形成 128 维的 SIFT 特征向量，对于每一张企业图像都包含多个关键点。

由于我们已经有了针对通用图像的识别算法，把它用于对企业相关图像的识别，具有较强的可行性。

6.3　本章小结

本章讨论了文本情感分析算法、短文本自然语言处理技术，并介绍了基于微博的文本内容的突发事件及全新突发事件预警方法

的应用。本章还研究了用户在紧急状态下的行为分析、社会网络推荐技术及基于图像内容的企业网络舆情管理的技术。

思考题

1. 简述短文本自然语言处理技术。

2. 掌握社会网络推荐技术及基于图像内容的企业网络舆情管理的技术。

3. 举例说明基于微博的文本内容的突发事件及全新突发事件预警方法的应用。

第7章

面向企业网络舆情的社会网络信誉及营销管理

本章学习目标

- 掌握基于内容和交易网络结构的信任测度。
- 熟悉基于企业网络舆情分析的企业网络营销管理方法。

7.1 面向企业网络舆情的社会网络信誉平台构建方法

企业信誉评价是企业品牌管理的组成部分(刘晓亮,2013)。信誉是通过虚拟组织的建立来实现的,解决方案是在网上建立信誉市场。开展信誉市场的第一步是所有的交易者都必须注册;第

二步通过客户向交易伙伴发送的反馈信息采用数据挖掘的方法(周晓飞和石勇,2009)来创建一个信用等级系统。

信誉系统的出现,使信誉变得数量化和可视化。作为一种非正式的、自律性的制度机制,信誉系统广泛使用于在线交易中。例如,易趣网和淘宝网都有社区信誉系统。在线信誉系统通过收集评价信息来计算客户的信任度。客户信任度将为交易客户提供极有价值的参照,为在线交易提供安全保障。在线信誉系统中,通常包括两个部分:信誉评价体系和反馈论坛。

一般地,网站对客户都有信誉评价纪录,客户在网站上交易成功后,在评价有效期内,双方就该交易互相作出信誉评价:好评(+1)、中评(0)、差评(-1)。众买方对某个卖方提供的评价累计后成为该卖方的信誉评价的重要指标,众卖方对某个买方提供的评价累计后成为该买方的信誉评价的重要指标。当买卖方达到一定的计算分值后,网站上会有一定的信誉(信用)等级图标出现,为买卖方提供交易决策支持。这样未来的交易者可以凭借信用等级系统估量对方的可信度。例如,在以往的研究中,Li 等(2008)通过采用多目标规划的方法对信用卡的持卡用户的信用水平进行分类和评估,李建平等和刘京礼等(2010)通过采用支持向量机及其改进方法结合信用卡的真实数据对持卡者的信誉进行分类和评估。

在信誉系统的体系构建上,一些学者建议:

(1) 引入社会征信体系,对买卖双方进行综合评级。对买卖双方的信誉等级评定不完全信赖交易的评价,而是与双方的社会

信用体系紧密联系起来，把好信誉评价机制的源头关。

（2）把物流行为与评价机制直接关联。对于有形产品的交易，除同城贸易以外，异地贸易一律与物流行为挂钩，没有物流行为的异地交易均视为虚假交易，以避免通过虚假交易“刷信誉”的行为。

（3）调整评价分值的范围和权重。把现行的三种分值改进为十分制或百分制，以增加评分的科学性。同时以每单贸易的产品价值或交易额为依据，设置不同的权重，增加高价格产品的权重，减少低价格产品的权重。

（4）根据产品的保修期，给客户重新评价的机会。

（5）完善恶评申诉体制，给卖方营造一个正常的成长环境。

作为主观评价方法的信誉系统，已经成为企业品牌管理中应用最广也是最重要的一种信誉管理方法。信誉系统保存了每个交易方的每一笔交易和相关的信誉度评价，这是陌生的交易双方之间建立信誉关系的基础。买方在交易前通过前人对卖方的评价来作出他的购买决策。同时买方对卖方的评价也会影响卖方以后的行为，他们会为了更高的信誉度表现出更好的行为。

建设社会网络的信誉系统平台，就是将各方信誉数据集中到一起，这样做的好处是，建立一个基于社会网络信誉的开发式的用户信用系统，便于任何需要它的用户或企业获取（在获得授权的条件下）。

7.2 基于内容和交易网络结构的信任测度

假设商品为 $P_1 \sim P_5$，买方为 $b_1 \sim b_4$，卖方为 $s_1 \sim s_5$，商品及买卖双方之间的关系网如图 7-1 所示。

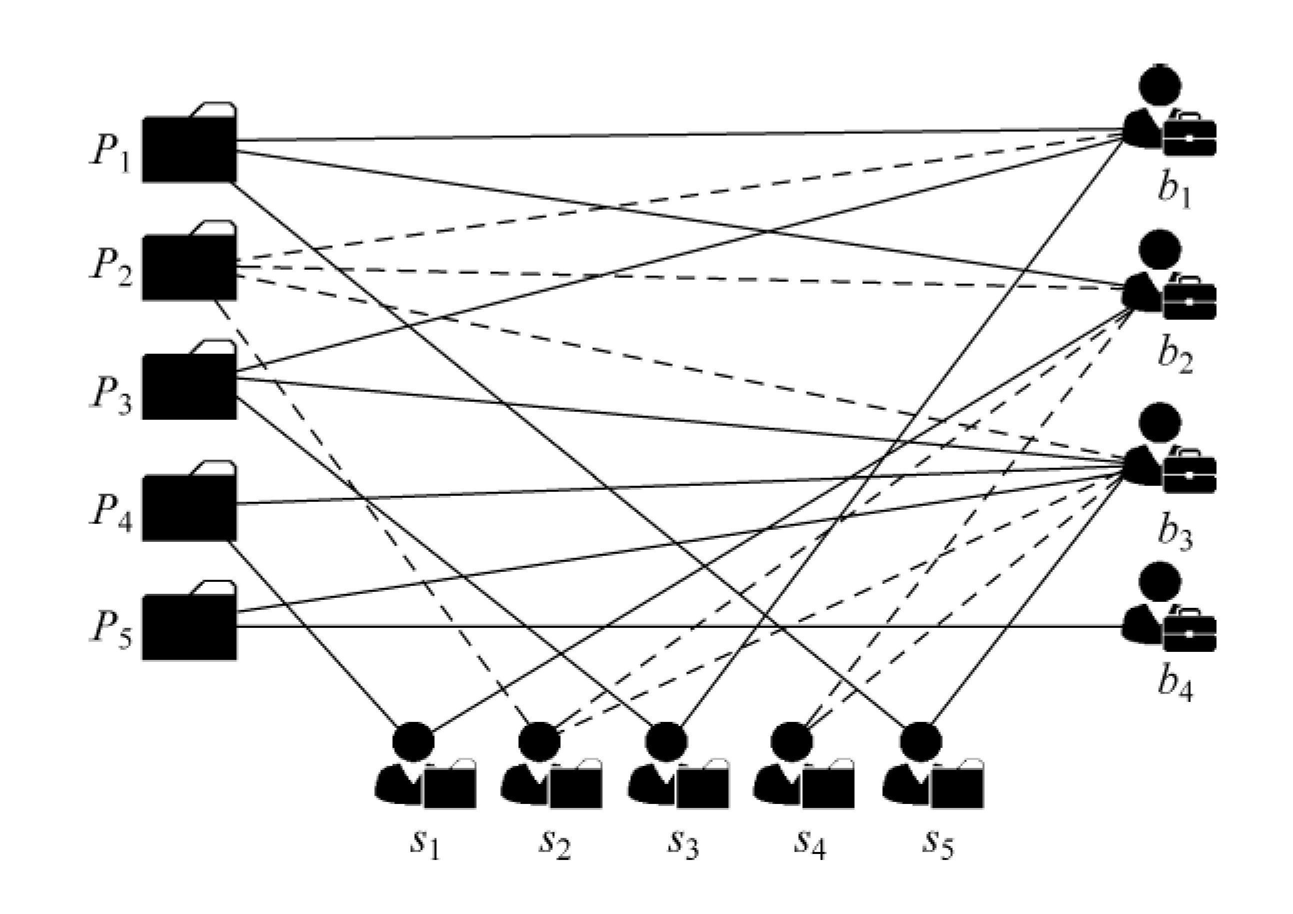


图 7-1 商品及买卖双方之间的关系网

设 b_i 和 b_j 有共同的卖方 s_k，则在买方-卖方交易关系 2 模式网中，我们可以计算 b_i 对 b_j 信任度。首先，我们定义一个基于交易和社会网络文本内容的混合信任度。如图 7-2 所示，w 是交易量，v 是客户褒贬值，我们定义一个向量，通过夹角反映信任度，$x_{ik}^{(n)}=(w_{ik}^{(n)},\overline{v_{ik}^{(n)}})$，其中 ${w_{ik}}^{(n)}$ 为 b_i 从 b_i 和 b_j 共同 s_k 购买次数，$\overline{v_{ik}^{(n)}}\in[-1,1]$为 b_i 对 s_k 给出的褒贬平均值，归一化(0,1)(时变的，不断修正)。

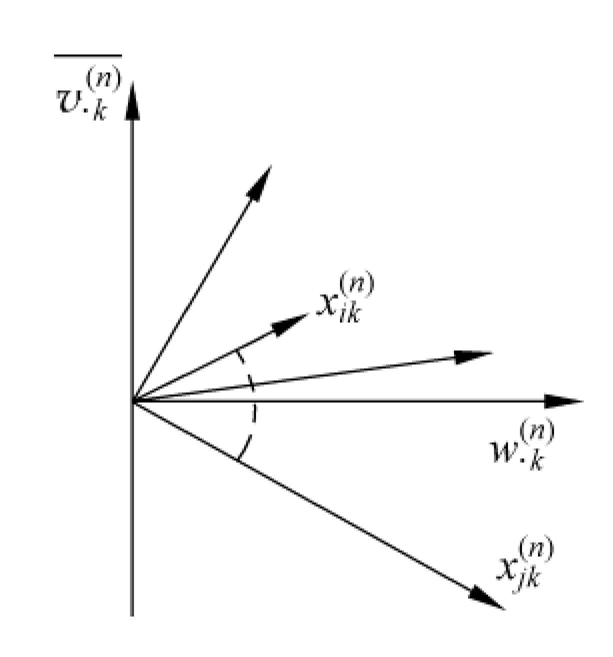


图 7-2 商品及买卖双方之间的关系网

显然,如果把更多的向量放在一起,可以多重比较,看哪两个客户相互信任度比较高。

类似地,我们可以通过(买方-商品关系 2 模式网)可得信任度,也可以得到加权平均值。

7.3 基于企业网络舆情分析的企业网络营销管理方法

网上购物带有着功利主义和享乐主义的情感(Machado,2005),传统的产品评论往往关注于把文本映射到给定的主题,如体育、经济、政治等(施国良和程楠楠,2011;李实,2012;Cantador,2011)。但是,对一个产品制造商而言,他们想知道互联网上那些"草根"客户对自己产品或者竞争对手产品的评价,因为"群众的眼睛是'雪亮'的"。了解"草根"客户对自己产品的评价,对他们自己的产品发展、市场和客户关系经营是非常有价值的。

基于社会网络舆情的营销与传统营销管理的区别都是为了满足用户的需求，但传统营销管理基本上还是以企业为中心，用户只能从用户生产出来的产品对其价值与质量进行关注与选择。而在社会网络时代，用户在整个销售过程中更起主导作用，以用户为中心的理念得到最大程度的体现，用户可以浏览商品，而且还可以设计产品，甚至可以决定商品的价格。

由此可见，基于社会网络舆情营销的新特点是：

（1）客户自发的信息共享汇聚，突破了企业单向的广告、推送等，一方面通过信任机制和模仿机制等带动了其他消费者的消费，另一方面也为企业个性化多样化创新制造和服务提供了机会。

（2）以客户价值为中心结成价值网络，改变了组织形式和演化机制（Jones 等，1997；Talluri 等，1999；肖渡和沈群红，2000；王伟，2005），每个企业不仅要构建或加入一个价值网络，而且要考虑自己在价值网络中的位置，避免低端锁定（胡大立，2006；杨瑞龙和冯健，2004）。

（3）这种信息技术支撑、信息资源共享的全球化商务模式具有更高的效率、更大的活力、敏捷性和创新性。

基于社会网络舆情的营销有以下功能和特征（Ko，2013）：

（1）购物列表通过搜集消费者的浏览、点击及购买数据，在消费者浏览商品时，自动推荐与该商品属性相似的产品。例如亚马逊、Best Buy、Kaboodle、Style Feeder、This Next、京东商城、天猫等。

（2）分享服务或产品允许人们通过文字、图片或视频等方式将购买商品或服务的过程、商品使用或享受服务的经验等通过社会网络分享给对此商品或服务感兴趣的消费者。是一种用于分享消费者的购买经历或享受服务的工具，从人们兴趣角度提供了社会网络佐证，例如蘑菇街、美丽说、辣妈酷宝、零食控。

（3）社会网络评论通过具有权威性质的专业评论者（微博评论、网络杂志评论、YouTube 评论等）和真实的消费者，赋予人民的声音以权威，让“草根”的力量显示出来。例如与其他购买者所提供的对于产品或服务质量的可信评价，最早出现的类似社会网络评论来自 Web 2.0 时代开启的 eBay 对卖家的评论。扩大到社会网络评论就加入了更多人的属性，不仅局限于交易阶段的评论，而是通过社会网络进行的评论。

（4）社会网络推荐系统通过对相似消费群体的挖掘提供个人推荐的工具，例如天猫、京东商城、亚马逊、当当等。

（5）推荐计划通过购物者信任的人们进行相关推荐。

（6）促销订阅获得那些大多数人不了解的独家促销。

（7）团购允许购买者成为一个大交易中的一份子的工具（Zhou 等，2013），而这些购买者在网上自动形成，彼此相互不认识。

（8）社交化网络店面针对社交化网络用户的独有的优惠促销、商品（Leitner 和 Grechenig，2008）。

所以，在 Web 2.0 环境下的企业营销活动更强调的是用户，重

点突出用户，以及这种社会网络活动给企业带来的增值。用户成为销售活动的中心，所有的活动依靠用户来进行，用户不再是作为被动的客体而是作为一种主体参与，用户除了作为使用者之外，同时成为主动的生产者、使用者和传播者。在这种新型的商务模式下，用户与商家的概念变得模糊，用户从简单的购买者变为全面整合的客户，甚至是营销者，他们不仅可以选购商品，还可以推销产品，在销售活动中获得利润。

7.4 本章小结

本章讨论了面向企业网络舆情的社会网络信誉平台构建方法和基于内容和交易网络结构的信任测度，提出了买方-卖方交易关系 2 模式网的夹角信任度方法，探讨了基于企业网络舆情分析的企业网络营销管理方法。

思考题

1. 掌握基于内容和交易网络结构的信任测度。
2. 举例说明基于企业网络舆情分析的企业网络营销管理。

第 8 章

面向企业社会网络舆情管理的用户行为理论

本章学习目标

- 掌握基于社会网络舆情内容的用户意图挖掘基础模型和基于用户关系网络拓扑及用户在信息传播中行为基础模型。
- 熟悉衍生模型。

8.1 引言

本章的研究力图回答这样一个问题：在社会网络大数据环境下，基于企业社会网络舆情的内容和结构的用户行为分析，及在此

基础上,对企业绩效的影响机制和企业管理新方法。

本章将对于企业社会网络舆情的主体(用户)的具体行为模式进行研究。通过对用户行为的研究得出社会网络大数据环境下企业舆情的具体静态(内容)以及动态(结构)的特点,为企业社会网络大数据环境下舆情管理方法的研究奠定基础。

我们先对社会网络中的用户行为给予形式化的定义。在给定时间(t_0,t_T)内,帖子p在网络中传播过程可以看作是一个网状结构,如图 8-1 所示。针对目前典型社会网络(例如微博),每一个用户(节点)对帖子可以采取的基本动作有发帖$a_0 \in [-1,+1]$、点赞$a_1 \in \{-1,0,1\}$、回复$a_2 \in [-1,+1]$、转发$a_3 \in [-1,+1]$、评论$a_4 \in [-1,+1]$、删除$a_5 \in \{0,1\}$,其中$\{-1,0,1\}$和$[-1,+1]$表示了转发、回复或评论帖子内容的情感倾向,负为贬、0 为中性、正为

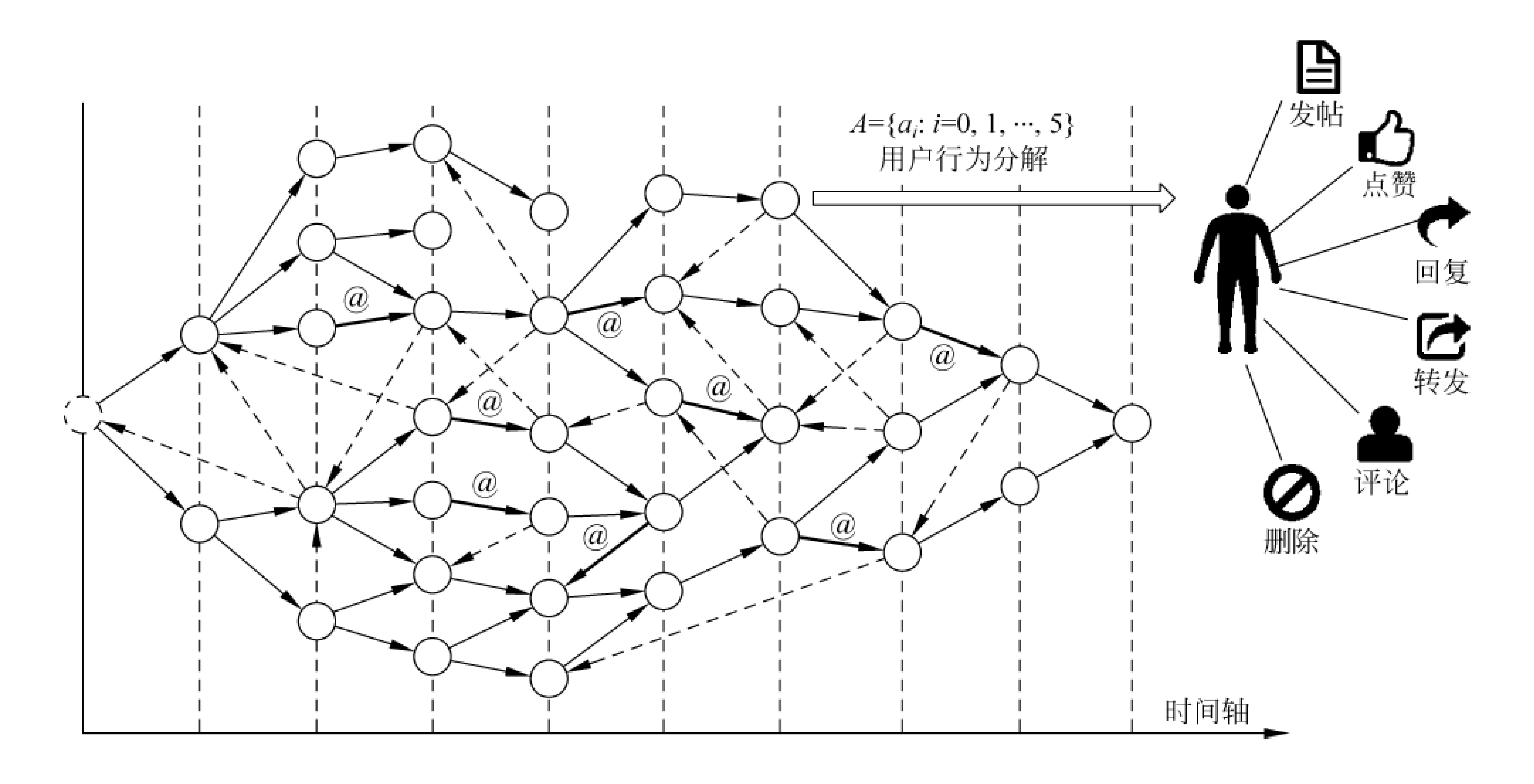


图 8-1　舆情传播过程的时间序列分解示意图

褒，{0，1}表示了是否删除，0 为否、1 为是。如果在转发和评论中有进一步的@动作，则相对于它在网络结构中对某些连线“加粗”。

令基本动作集为 $A=\{a_i: i=0,1,\cdots,5\}$，其中动作 a_i 有五个属性：主体 s、帖子 p、时间 t、地理位置 l、对象 o 和其他特别属性 b，定义动作 $a_i(s,p,t,l,o,b)$，其中主体 s 表示用户，地理位置 l 表示了包含使用移动终端在内的用户发帖时所处的地理位置，对象 o 是指主体发出的帖子 p 所到达的节点。

网状结构中的用户动作可以用如下序列表示，$L_p=\{a_i(*,p,t_0,*,*,*),a_i(*,p,t_1,*,*,*),\cdots,a_i(*,p,t_T,*,*,*)\mid a_i(*,p,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$。于是，帖子 p 在网络中传播的节点个次(因为可能一个节点两次或以上)为 $|L_p|$。

设在给定时间 (t_0,t_T) 内，网络中传播的信息集合为 $P=\{p_0,p_1,\cdots,p_N\}$，则网络中的所有动作可以用 $L_P=\{L_{p0},L_{p1},\cdots,L_{pN}\}$ 表达，总数为 $|L_P|$。

同理，在给定时间 (t_0,t_T) 内，网络中的某个用户 s 的行为也可以用如下序列表示，$L_s=\{a_i(s,*,t_0,*,*,*),a_i(s,*,t_1,*,*,*),\cdots,a_i(s,*,t_T,*,*,*)\mid a_i(s,*,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$。于是，用户 s 在时间 (t_0,t_T) 内的行为的总数是 $|L_s|$。同理可以得出其他的序列 L_l、L_o 及其数目。

如果设在给定时间 (t_0,t_T) 内，网络中参与传播的用户集合为 $S=\{s_0,s_1,\cdots,s_M\}$，则网络中的参与用户的行为集合可以用 $L_S=\{L_{s0},L_{s1},\cdots,L_{sM}\}$ 表示，其总数为 $|L_S|$。显然，应该有 $|L_P|=|L_S|$。

本章把模型分为基础模型和衍生模型。为了清晰，基础模型编号为罗马数字，衍生模型编号为英文字母。

8.2 基础模型

在本部分中，基于社会网络计算，我们将提出两个面向企业社会网络舆情的用户行为基础模型，分别是：基于社会网络内容的用户意图挖掘模型（模型Ⅰ）、基于社会网络用户之间关注或“粉”行为静态结构及用户之间舆情动态传播行为的模型（模型Ⅱ）。其中，模型Ⅰ是模型Ⅱ的基础。模型Ⅰ是对用户帖子内容的挖掘，主要涉及发帖 a_0、回复 a_2 及转发 a_3 等。模型Ⅱ是对网络结构的分析，主要涉及点赞 a_1、回复 a_2、转发 a_3 及评论 a_4 等。

1. 基础模型Ⅰ（基于社会网络舆情内容的用户意图挖掘模型）

基础模型Ⅰ主要用于对企业社会网络中舆情的静态属性（文本内容和图像内容）进行研究与建模。

（1）基础模型Ⅰ-1（基于社会网络文本内容的用户情感倾向挖掘模型）。

在本章中，我们也选用其成熟的情感倾向分析技术，用于社会网络下的企业舆情研究中。

我们把社会网络中与企业相关的情感倾向值定义为企业社会网络舆情情感倾向值，简称企业舆情值。

由于我们的研究是针对企业的，所以我们需要建立一个知名

企业名单库，通过社会网络手工信息收集等多种途径进行材料的收集，将知名企业名单纳入名单库。在已有的研究中，对每一个存储在库中的待研究企业，我们已经收集了其各类简称、上市股票代码、企业领导人姓名等，但主要目的是为了进行针对其股市波动的研究。我们将挑选一个先期和我们有良好关系的典型企业，进行案例研究。

特别需要指出的是：对微博等社会网络来说，信息多以短文本的形式出现。所以，研究短文本条件下的企业舆情也非常重要。进一步，有效地实时监控这些在线文本也是非常重要的。因为在线文本中恶意的负面“流言”可能导致企业信誉严重流失，及时发现并加入正面的信息，就可以有效地防止问题的发生，从而反过来正面引导和影响社会网络的企业舆情，如图 8-2 所示。所以，上述应用相当于“企业级”的舆情分析监控。事实上，为了完成国家层面上的舆情管理（陈华和梁循，2007），相关部门早就建立了国家级的舆情监控中心。

（2）基础模型Ⅰ-2（基于社交网络图像内容的特定用户意图分析模型）。

显然，包含图片的微博携带的信息量更大，研究发现包含图片的微博有明显多的转发量。

对企业商品或服务的评述，有的配以图像，其中一部分图像是经过手机拍摄，通过无线网络直接传播到微博上的。如何在庞大的数据中甄别与提取特定企业相关图像，并进行深入的分析，一直

图 8-2 企业舆情情感倾向值示意图

是一个比较难的课题。显然如果泛泛地做难以出成果，所以，在具体的研究中，可以挑选一个特定的主题（例如某个企业的突发事件），针对该特定企业展开研究。我们研究的内容缩小到检测发现与某特定企业相关的图片。首先，我们可以收集该企业的相关图片，建立其图像库。图像库包括企业高管照片、企业 logo 及中英文名称等。其次，使用 SIFT 等算法对图像特征进行标注，并进一步利用 SVM 等机器学习方法进行分类学习。最后，将机器学习结果应用到对企业社会网络舆情突发事件的预警中去。

2. 基础模型Ⅱ（基于社会网络结构的模型：用户关系网络拓扑及用户在信息传播中的行为）

基础模型Ⅱ主要用于对企业社会网络的动态属性（网络传播结构）进行研究。通过这部分的研究挖掘企业的舆情动态传播特性，为社会网络复杂环境下企业舆情的管理方法的研究奠定动态控制的基础。对社会网络结构的研究，本模型不仅涵盖网络社区

的发现与分析，还将探讨关键路径和权威用户，也就是说，本模型将分别从“面”、“线”和“点”来进行。

（1）基础模型Ⅱ-1（用户中的社区发现：用户群体行为模型）。

社会网络多样化的另一个突出的特征是异质网络的出现，异质网络指的是网络中存在多种类型的用户和多种连接关系，用户的类型通常表现为内容，用户之间的关系是反映了用户之间的联系。现阶段绝大多数的用户社区发现算法往往将用户联系同用户内容相隔离，从而导致其社区发现结果不够合理，而少数综合用户联系和内容的用户社区发现算法较为复杂；LCA 算法是社区发现算法中算法效率较高，且社区质量较好的算法，然而，其在聚类时未考虑边的真实兴趣体现。

针对这些问题，本章也将构建配属了关注关系的加权生活网络，以关注关系之间是否有共同用户为关注关系潜在的边，以关注关系所关联用户的兴趣集的交集为关注关系的兴趣特征，我们构建了微博网络 R-C 模型（周小平和梁循，2014），如图 8-3 所示，其中，U_i、T_i 和 L_i 分别为传统微博模型中的用户、微博和用户关系；R_i、C_i 和 U_i 分别为微博网络 R-C 模型中的用户关系、用户关系特征和用户关系间潜在的连接。在此基础上，我们可以研究这些社区结构和特定企业舆情管理上的作用，并给企业管理提供指导。社区发现的结果可以用于进行对企业有利信息的精准投放或精准的个性化推荐，为企业舆情调控提供支持。

社区发现模型其实可以有很多种。我们观察到，上述社区发

图 8-3　社会网络 R-C 模型图

现模型，实际上只是对物理存在的拓扑网络结构进行的社区发现。如果把用户的内容属性考虑进去，会进一步提高模型的实用性。所以，我们也计划进一步研究基于内容的某领域（例如足球、育儿）的兴趣社区发现模型，或称为共现词社区发现模型。

具体地说，首先从社会网络 G 中，抽出一个的兴趣子网（例如妈妈育儿子网）或特定子网，记为 G_1，然后在这个兴趣子网 G_1 中进行社区发现。此外，上述特定子网为面向某企业的子网。可以看出，子网也是基于社会网络拓扑结构的模型。

在本章研究中，为了叙述简单，我们在 G 和 G_1 都适用的情况下，只简单地写为 G。

（2）基础模型Ⅱ-2（用户网络中的路径：用户动态行为模型）。

针对企业的舆情，识别舆情传播的路径至关重要，尤其是这些传播路径中的关键路径。显然，对某条信息的传播来讲，处于关键路径中的用户，其影响力比其他用户要大。

关键路径分为两大类，基于网络拓扑结构的静态关键路径和基于动态传播的关键路径。其中，基于网络拓扑结构的又分为社区内关键路径、跨社区关键路径，如图 8-4 所示。在一个社区中，两个点之间的最短路径，称为社区内关键路径。

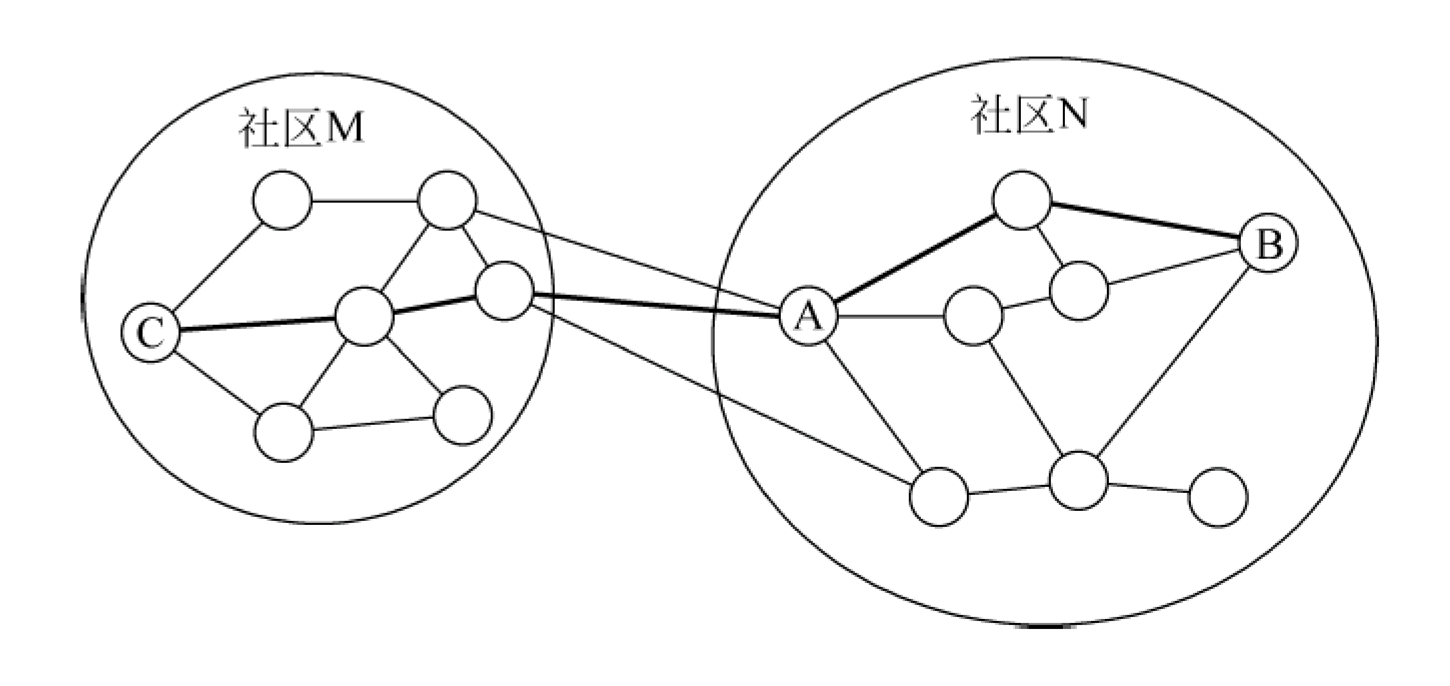


图 8-4　基于网络拓扑结构的关键路径发现

首先我们将两个相邻用户之间的连线称为线段。在跨两个社区的所有连线中，出进这两个社区的用户的度的总和最大的那条线段，称为两个社区之间的桥。对于任何跨社区的两个用户，经由桥的最短路径称为它们之间的跨社区关键路径。

基于动态传播关键路径的分析，需要考虑传播信息最多的路径，它也分为社区内关键路径、跨社区关键路径。

在信息实际传播中，不同用户，转发和评论信息量不同，有的用户是喜欢转发和评论的“大嘴”，有的喜欢“潜水”。在社会网络中，假设两个用户一个在北京，一个在广州，只要它们和用户 i 是相邻用户，信息从用户 i 到达两个用户的时间就是相同的。

在一个社区 G 内，在某个单位时间内，用户 i 转发和评论信息

量为 W_i，信息从用户 i 传播到用户 j。这两个用户间的实际路径可能有多条，设第 k 条路径上的用户为 $i, i_{k_1}, \cdots, i_{k_n}, j$，则路径 k 上的所有用户转发和评论的总信息量为 $W_i + \sum_{m=1}^{n} W_{k_m}$。在用户 i 和 j 之间的所有路径中，称第 $l = \arg\max_k \sum_{m=1}^{n} W_{k_m}$ 条路径为关键路径（"大嘴"传播路径）。在实证时，在给定时间 (t_0, t_T) 内，一个帖子 p 在网络中传播过程可以看作是一个网状结构，网状结构中的用户动作可以用如下序列表示，$L_p = \{a_i(*, p, t_0, *, *, *), a_i(*, p, t_1, *, *, *), \cdots, a_i(*, p, t_T, *, *, *) \mid a_i(*, p, t, *, *, *) \in A, i \in \{0, 1, \cdots, 5\}\}$。记 $n_p = |L_p|_{i=3,4}|$。关键路径（"大嘴"传播路径）就是找最大的 n_p 对应路径。

在跨两个社区的所有连线中，在某个单位时间内，出进这两个社区的用户被转发和评论的总和最大的那条线段，称为两个社区之间基于内容的桥。对于任何跨社区的两个用户，经由基于内容的桥的最短路径称为它们之间的跨社区基于内容的关键路径。

在社会网络中，用户是否去发帖 $a_0 \in [-1, +1]$、点赞 $a_1 \in \{-1, 0, 1\}$、回复 $a_2 \in [-1, +1]$、转发 $a_3 \in [-1, +1]$、评论 $a_4 \in [-1, +1]$、删除 $a_5 \in \{0, 1\}$，构成了用户的"动态"行为。由于用户在网络中所处的不同社区、路径以及自身用户的重要性等，都对企业舆情产生了不同程度的影响。

当然，关键路径还可以有更复杂的定义，因为影响关键路径的因素很多，这可能涉及更复杂的 $\{a_i(*, p, t_0, *, *, *),$

$a_i(*,p,t_1,*,*,*),\cdots,a_i(*,p,t_T,*,*,*)|a_i(*,p,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$表示。事实上,在网络中,信息传播不仅是和具体内容以及会和时间等因素有关,而且信息在传播过程中还会衰减。如果将这些信息考虑进入,会使得路径的描述更有意义。

对任何两个不直接连接的用户之间,可能存在多条路径,这些路径可以形成串联和并联方式,如图 8-5 所示。在图 8-5 中,b 代表网络中的用户,该用户可能含有很多内容信息和我们前面模型计算出的情感倾向值、图像内容标签等(曹润和梁循,2012;施晓菁和梁循,2012)。

对串联方式来说,假设某一条路径上用户总数为 $N+1$(含首尾用户),第 0 个用户(即首用户)的信息值 W_0,则第 N 个用户(即尾用户)的信息值为 $\prod_{n=1}^{N} x^{-n}W_0$。

对并联方式来说,设 M 为到达用户 b_i 的路径数(例如在图 8-5 中,$M=5$),则用户 b_i 的信息值为 $\sum_{m=1}^{M} W_m$。

除了串联和并联方式外,信息在传播中还存在着衰减和放大情况。在本章中,我们可以研究和实验各种内容信息衰减率的规律。

首先,路径越长,信息衰减的可能性越大。我们设任意两个相邻用户的距离为 1,k 为传播的用户的个数(即用户与信息源用户的距离)。我们可以假设内容信息衰减方式是 x^{-k},其中 x 是一个大于等于 1 的参数,它需要通过实证确定,对信息源,$x^{-k}=1$。如

果我们取 $x=e$，则对第 1 个用户，$x^{-k}=0.37$，…，以此类推。

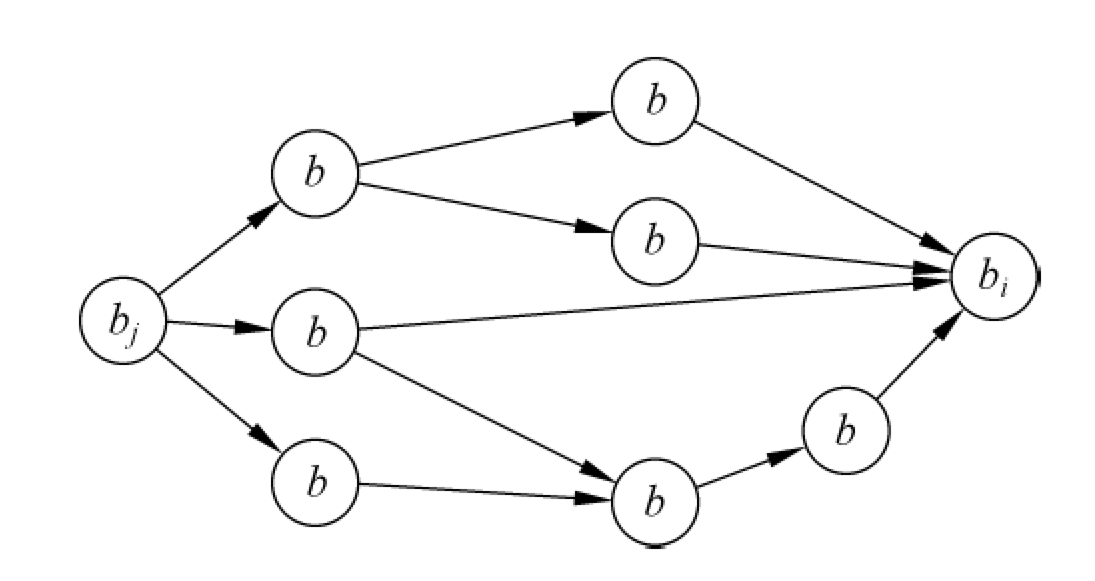


图 8-5　社交网络中内容信息传播路径的串联和并联方式

其次，内容增强或减弱也会影响信息强度。在上述的 x^{-k} 前面，我们定义一个强度转换系数(乘子)a_{ij}，在用户 i，如果用户只是转发了相邻用户 j 的信息，则设 $a_{ij}=1$，如果用户给了褒义的评论则设 $a_{ij}>1$，如果用户给了贬义的评论则设 $a_{ij}<1$，其中 a_{ij} 的大小根据实证结果确定。

(3) 基础模型Ⅱ-3(权威用户：用户影响行为模型)。

在社会网络中，用户所在网络中所处的位置和周边结构的不同，决定了它们具有不同的角色意义，如图 8-6 所示。那么不同位置、度数不同的用户所掌握、控制的资源能力和数量会有着巨大的差异。比如说，有些用户处于社交网络中的核心位置，有些用户却处于社交网络的边缘位置，还有些用户则在社交网络中扮演着“桥”的角色。此外，用户的影响力也不同，有些影响力大，有些影响力小。因此，对于企业的突发事件舆情信息管理来说，区分社交网络中的各种用户至关重要。

根据社会网络的传播特性，尤其是微博的出现和普及，不仅使

得社会网络中的信息能够在较短时间内实现较大范围的传播，也打破了传统的以政府为主体的“单级”信息传播控制模式。

本模型给出 4 个企业舆情权威用户的定义，并提出了在海量社交网络中识别舆情权威用户的模型。

第 1 个模型是：在一个社区中，具有最大（k 个）度的用户，称为权威用户。我们以用户最大的度为例说明。我们定义用户 i 的度 d_i 为与该用户直接连接的用户数目，则第 $\operatorname{argmax}_i d_i$ 个用户为权威用户。在去掉第 1 个权威用户后，第 2 个权威用户可以类似得到，依次类推，可以找到前 k 个权威用户。这是在社会网络中一般的定义方法。

第 2 个模型是：在某个单位时间（t_0，t_T）内，如果某个社区中的某（k 个）用户发布的信息最多，称为权威用户。我们以某用户发布的信息最多的用户为例说明。设用户 i 发布信息数为 W_i，则第 $\operatorname{argmax}_i W_i$ 个用户为权威用户。在实证时，使用计算机很容易完成上述统计工作，即如果在 $L_s=\{a_i(s,*,t_0,*,*,*),a_i(s,*,t_1,*,*,*),\cdots,a_i(s,*,t_T,*,*,*)\mid a_i(s,*,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$中，记 $n_s=|L_s|_{i=0}|$。如果与其他主体相比，n_s 的最大，则称 s 为权威用户。依次类推，可以找到前 k 个权威用户。

第 3 个模型是：在某个单位时间内，如果某个社区中的某（k 个）用户发布的信息被回复、转发或评论得最多，称为权威用户。同前面类似，设用户 i 发布的信息被回复、转发或评论得最多，其总数为 W_i，则第 $\operatorname{argmax}_i W_i$ 个用户为权威用户。在实证时，如果

在 $L_s=\{a_i(s,*,t_0,*,*,*),a_i(s,*,t_1,*,*,*),\cdots,a_i(s,*,t_T,*,*,*)\mid a_i(s,*,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$ 中，记 $n_s=|L_s|_{i=2,3,4}|$。如果与其他主体相比，n_s 的最大，则称 s 为权威用户。依次类推，可以找到前 k 个权威用户。

第 4 个模型是基于回复、转发或评论内容的，但涉及情感倾向，设用户 i 发布的第 j 条信息被回复、转发或评论的情感倾向值为 W_{ij}。于是，对用户 i 以外的所有用户 j，经过衰减作用 a_{ij} 后，用户 i 发布的第 j 条信息的实际影响强度为 $\sum_j a_{ij}W_{ij}$，我们定义用户 $\mathrm{argmax}_i \sum_j a_{ij}W_{ij}$ 为权威用户。在实证时，如果在 $L_s=\{a_i(s,*,t_0,*,*,*),a_i(s,*,t_1,*,*,*),\cdots,a_i(s,*,t_T,*,*,*)\mid a_i(s,*,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$ 中，记 $n_s=\sum_{t_0}^{t_T}(a_2+a_3+a_4)$。如果与其他主体相比，$n_s$ 的最大，则称 s 为权威用户。依次类推找到前 k 个权威用户。

显见，第 1 种定义是纯基于网络结构的，是后两者的基础。一般地，在一个社区中一个度不是很大的用户，很难成为被回复、转发或评论数目最高的几个用户。第 2、3 种是基于数量的，第 4 种是基于内容的。其中第 3、4 种要求比较严格，即如果要认为某用户是权威用户，其粉丝必须明确“表态”，即回复、转发或评论。

在确定了权威用户 m 后，我们可以研究：给定一个用户 b，如何在多条串联、并联路径中，确定信息由 m 至 b 的关键路径以及信息传播到 b 时的信息衰减程度，研究 m 对 b 的影响力，如图 8-6 所示。

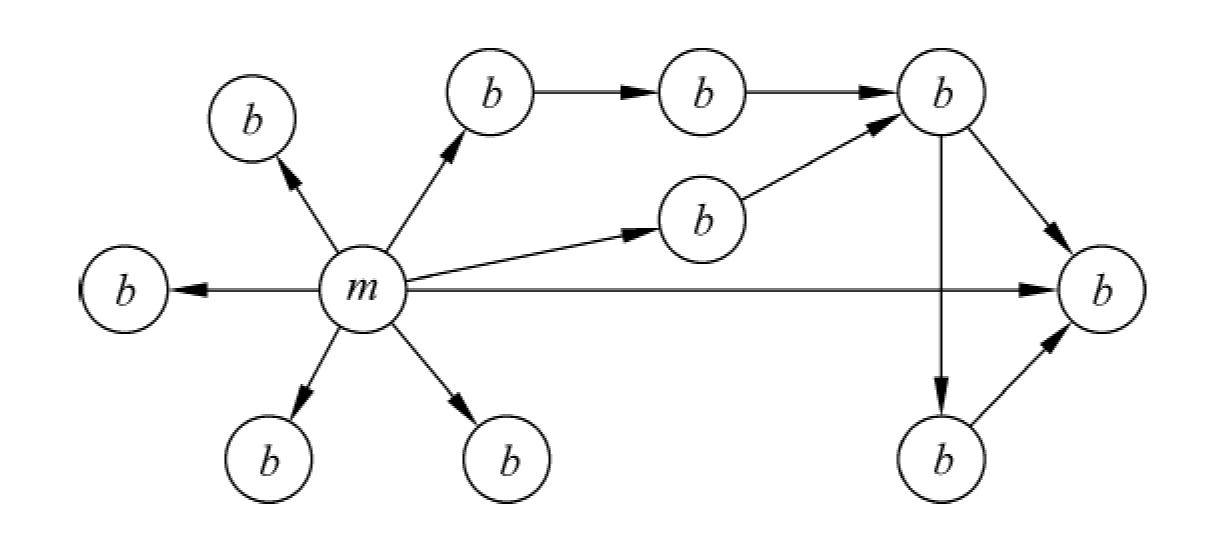


图 8-6　权威用户 m

8.3　衍生模型

上述基础模型主要是针对社会网络的本质特征和用户基础行为的分析，进一步，我们可以把基础模型延伸至针对企业舆情分析的特定主题中，提出更加具体可行、针对性更强的衍生模型。

1. 衍生模型 A（基于社会网络金融信息情感倾向值与股市波动关联分析）

基于内容分析的模型Ⅰ-1，我们可以研究社会网络金融信息的情感倾向值与上市公司股票价格的影响机制。通过收集社会网络上的信息，观察评论的情感取向对企业是否能够提供满足用户需要的产品和服务做出测评，并对企业管理中的业务流程、成本结构、销售策略提出新的建议。

股价的波动率一般认为是在某一个时间段内的价格的方差，它的大小往往意味着不确定性和风险的大小。如果我们能够对波动率进行预测，则对于企业及其投资者有着重要的意义。由于金融市场的波动率可以表现为一个时间序列，其当前的值与其之前

的一系列值以及外界变量有着密切的关系，我们利用 SVM，使用信息情感的倾向值（其中褒为正值，贬为负值，大小表示褒贬强度），探讨金融信息情感倾向与金融市场波动率的非线性关系。

对全网的舆情分析，我们可以研究对象细化到某几个典型的上市企业，并将舆情分析的数据目标改为社会网络数据。具体地说，首先，针对某企业建立一个用户词典，并在金融舆情信息中，过滤出相应的上市公司信息。随后，可以引入社会网络的情感信息，定义 W 为基于社会网络的基于模型Ⅰ-1 的情感倾向值。针对该企业，我们将重点研究 W 的分析算法。本模型主要涉及用户 a_0、a_1、a_2、a_3、a_4 相关帖子的内容。W 的分析算法的目的就是建立一个社会网络帖子的文本信息与 W 的关系。我们的具体做法是，通过已有的成熟的自然语言处理技术，确定某条股市信息是利好还是利空，并从用词中确定其量度大小，并根据其他辅助信息例如信息长度，得其强度 W。具体地说，令在时间 t，在社会网络上有关该企业的第 l 条股市信息的强度为 $W_{t,l}\in[-1,+1]$。对于利多信息，令 $W_{t,l}>0$；对于利空信息，令 $W_{t,l}<0$。利多利空的强度由 $|W_{t,l}|$ 确定，$|W_{t,l}|$ 越大，认为利多利空的强度越大。记 $W_t=\sum_{l} W_{t,l}$。于是我们得针对该企业的网络金融信息流时间序列。

实验表明，该时间序列具有一定的季节性，以 7 天为一个周期，所以我们做 7 阶差分消除季节性。我们还观察到，常常地在一段时间内，金融信息流时间序列变化相对小；而后变化又相对大。

上面的模型可以给我们提供一个“政策实验室”，研究该企业

的绩效和股价的关系。这意味着如果该企业的绩效好，股价就会成上涨趋势，否则呈下降趋势。企业可以利用社会网络的金融信息时间序列，适当增加利好信息，减少利空信息，从而影响在线舆情。

2. 衍生模型B(基于微博文本内容的全新突发事件的用户反应)

本模型基于内容分析的模型Ⅰ-1，研究基于微博的全新突发事件发现问题。由于对常规的反复出现的突发事件的用户反应，学者们已经提出了很多研究模型。具体研究时，可以不求全面，只研究全新突发事件的用户反应及相应的企业舆情管理方法。

监控企业舆情的全新突发事件，我们只要把《突发事件字典》替换成《企业突发事件字典》即可。在现实中，对特定企业，也会出现完全意想不到的突发事件，某类信息，在微博上开始迅速传播。对计算机文本分析技术来说，上述突发事件，一般是不好放入事先给定的用户词典中的。但如果没有相应技术，企业舆情管理就是一个只会识别过去发生过的类似事件，不能对全新事件“举一反三”进行监控的“傻孩子”。

本模型可以研究用户在企业出现全新的突发事件情况下的反应，提供给企业管理者进行决策支持。它通过实时监控并处理微博文本信息，可以自动及时发现社交网络中包含的全新企业突发事件，并及时利用社会网络，增加正面的有引导性的信息，通过其他手段联系或删除负面的信息，从而借助积极的操作，影响在线舆情，使舆情向着对该企业有利的方向发展。对出现负面的信息，我

们可以研究企业应该如何利用在线舆情进行服务挽回等管理措施，并进一步研究这些措施对股市的影响机制。

由于对微博等社会网络大多以短文本的形式传播，我们可以专门研究短文本条件下的企业突发事件的舆情管理研究。

再进一步，使模型B结合模型A，也可以成为研究社会网络与上市公司股价关系的重要手段。当企业出现全新突发事件时，该企业的股价都会出现相对比较剧烈的波动。

3. 衍生模型C（用户口碑主导网络中企业舆情管理优化）

在网络拓扑结构上，它的特点是，粉丝多，但被“粉”也同样多，如图8-7所示。这类网络可以分为陌生人购物分享网络，例如蘑菇街、美丽说，以及熟人圈社交网络，例如微博、人人网等。本模型主要涉及用户的所有动作$a_0 \sim a_5$。在用户口碑主导网络中，没有明显的权威用户存在，其特定是没有度突出大的节点，很大比例的节点“影响力”比较均衡。我们可以进一步理清并给出严格定义。显见，它基于模型Ⅱ-2。

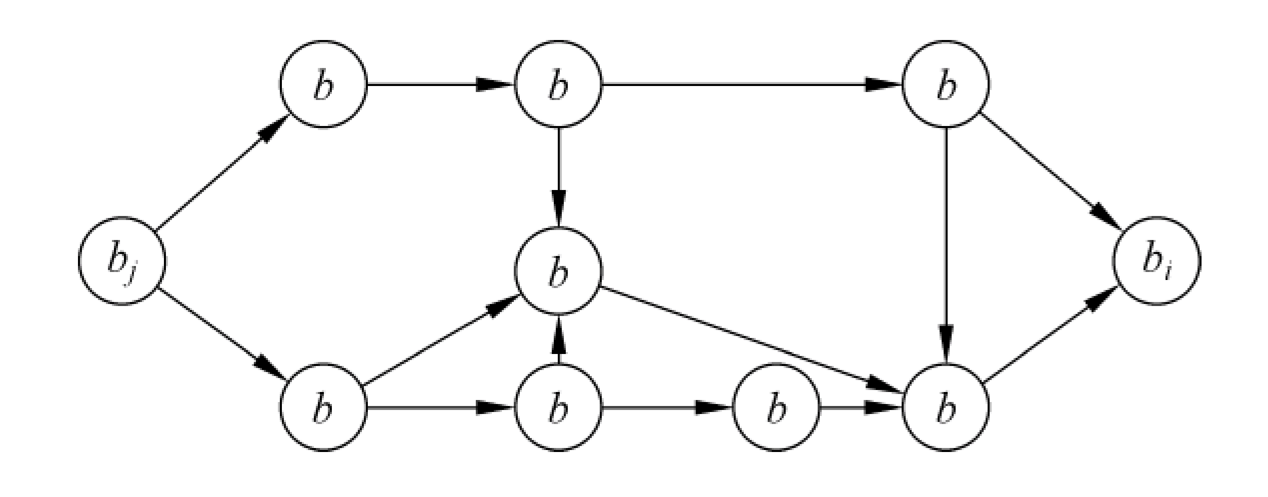


图8-7　口碑主导网络

在企业舆情优化管理上，这样的研究探索可以用于信息传播最大化和广告效果的计量，并可计算出在社会网络中选取哪些用

户播放对企业最有利的信息，可以使信息得到最大化的传播，在广告投入资金固定的情况下，可以帮助企业获得最大的收益，从而改善了营销管理。

我们可以研究怎么在这类网络中，以较小的成本，完成舆情管理任务。这实际上是一个企业舆情管理的优化问题。具体地说，首先，管理者在这类网络中，基于模型Ⅰ-1，找出针对某类信息传播的关键路径，并对关键路径上的节点用户进行“公关”，建立友好关系、增加正面信息，经常转发或评述与企业直接和间接相关的信息，扩大企业品牌效应，而对于企业的负面消息，力争使之少进行或不进行扩散，或使之进行正面引导，从而使得企业舆情对该企业有利，从而做得“事半功倍”的效果。其次，我们可以研究社区间的“桥”，并对其用户进行“公关”，建立友好关系，为企业舆情优化管理范围。

4. 衍生模型D(社会网络瘾分析)

我们可以使用管理科学中定量手段，结合心理学研究的特点，研究网络中用户 s 有社会网络(例如微博)瘾的行为特征。借助计算机，在社会网络的较大规模数据范围内，分地理区域、时间段等，发现相应的“瘾”用户，同时对具备社会网络瘾特征的用户行为特点进行定性和定量的分析，得出这类用户在企业社会舆情传播过程中的作用特点，从心理学模型角度得出对这类用户的干预和管理策略，并探讨在企业舆情管理中的应用。

首先，我们定义什么是社会网络瘾。

定义(社会网络瘾):在给定时间(t_0,t_T)内,设用户 s 在某社会网络中的行为序列为$L_s=\{a_i(s,*,t_0,*,*,*),a_i(s,*,t_1,*,*,*),\cdots,a_i(s,*,t_T,*,*,*)\mid a_i(s,*,t,*,*,*)\in A,i\in\{0,1,\cdots,5\}\}$,于是,用户 s 在时间(t_0,t_T)内的行为的总数是$|L_s|$。如果$|L_s|>\eta$,其中 η 为事先给定的常数,则称用户 s 有社会网络瘾。

显见,确定 η 可以有多种,例如 η 与时间段长度 t_T-t_0 有关。我们可以进一步研究并给出严格定义。

如果 L_s 中对 a_0 的数目大于 η_0,其中 η_0 为事先给定常数,则认为在该时间段,用户 s 是“发帖”较多。

其次,我们可以在较大规模的数据范围内,分地理区域、时间段等,分别统计并研究有社会网络瘾用户的比例和分布情况,得出用户的行为特征,制订相应的管理策略。

本模型也可以进一步扩展到研究对用户更细化的情形,例如,不同年龄、职业的用户的社会网络瘾情况及其行为特征。研究企业在舆情管理中对社会网络瘾情况的针对性管理措施。

5. 衍生模型 E(根据用户对特定事件的反应发现用户兴趣社区)

我们设想,一个能引起一类人群强烈“共鸣”的突发事件,应该首先在其社区内广泛传播,到一定时间后,才广泛传播到社区外。例如,婴儿奶粉忽然查出了问题,年轻妈妈群体(社区)会首先做出迅速反映。所以,对 G 包括兴趣子网 G_1 的社区发现,都只是一个静态的社区发现,并没有反映出一个动态的信息“传播

速度”问题。曹润(2013)通过实证,已经完成了一个对新浪微博的预研。

研究企业舆情的动态传播规律,对企业舆情管理有重要意义。在企业出现了不可控的突发事件后,对该企业感兴趣的群体会迅速在社会网络上回复、转发或评论这个突发事件。如果企业能够迅速发现并锁定含信息源的社区及其连接外部的桥(见模型Ⅱ-3),则可以通过公关办法,“断掉”这个桥(因为企业不可能断掉所有的用户之间的连线),从而以较低成本、大规模地减低信息传播的速度和范围。

8.4　本章小结

本章是本书的核心,重点讨论了基于社会网络结构的一系列模型。具体地说,本章提出了两个基于社会网络结构的基础模型,即基于社会网络舆情内容的用户意图挖掘模型和基于用户关系网络拓扑及用户在信息传播中行为的模型。在此基础上,本章研究了五个衍生模型,包括基于社会网络金融信息情感倾向值与股市波动关联分析模型、基于微博文本内容的全新突发事件的用户反应模型、用户口碑主导网络中企业舆情管理优化模型、社会网络瘾分析方法及根据用户对特定事件的反应发现用户兴趣社区方法。

思考题

1. 掌握基于社会网络舆情内容的用户意图挖掘模型。

2. 掌握基于用户关系网络拓扑及用户在信息传播中行为的模型。

3. 熟悉基于微博文本内容的全新突发事件的用户反应等衍生模型。

4. 了解根据用户对特定事件的反应发现用户兴趣社区方法。

第9章

社会网络舆情大数据的分解算法

本章学习目标

- 了解问题的环境和解决问题的思路及框架。
- 熟悉社会网络大数据的分解模型。
- 了解大数据领域的发展趋势。

9.1 问题的环境和解决问题的思路及框架

在实际中,社会网络舆情的数据量非常大,所用计算机的内存容量常常按维数平方增加,所用计算机机时常常按维数立方增加。

如果能将这些数据按存储地点、存储方式分解成若干个“子问题”,“子问题”在相同的或不同的“云”上。先解决这些子问题的优化和决策问题,每一个子问题相当于一个“智能代理”(intelligent

agent,IA),由这些 IA 自行解决各自的优化和问题,然后再综合考虑它们之间的关联,进而解决总体上的优化和决策问题。也就是说,上级 IA 给出指令,下级 IA 完成子问题的优化决策。这样总体优化和决策问题,就变得相对容易些了。

事实上,解决总体问题的“上级”也可以看作一个 IA,只是这个 IA 的输入是下级的优化结果,输出是对下级的指令。

下级 IA 完成的这些子问题的优化决策问题,既可以是数值型数据,也可以是文本型数据,还可以是图像、音频、音像数据。下级 IA 还可以有下级 IA,如图 9-1 和图 9-2 所示。

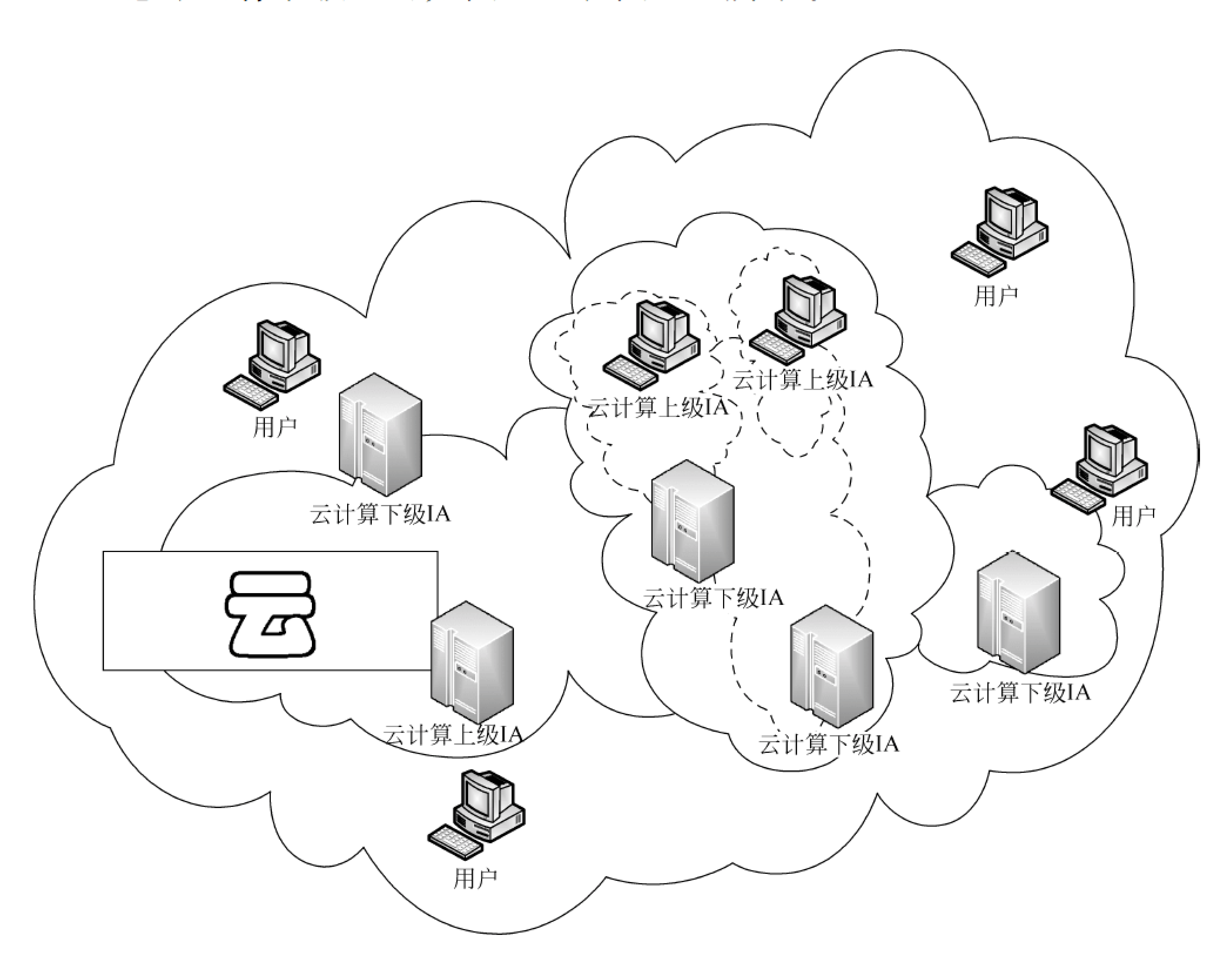


图 9-1　大数据系统的分布式云计算思路

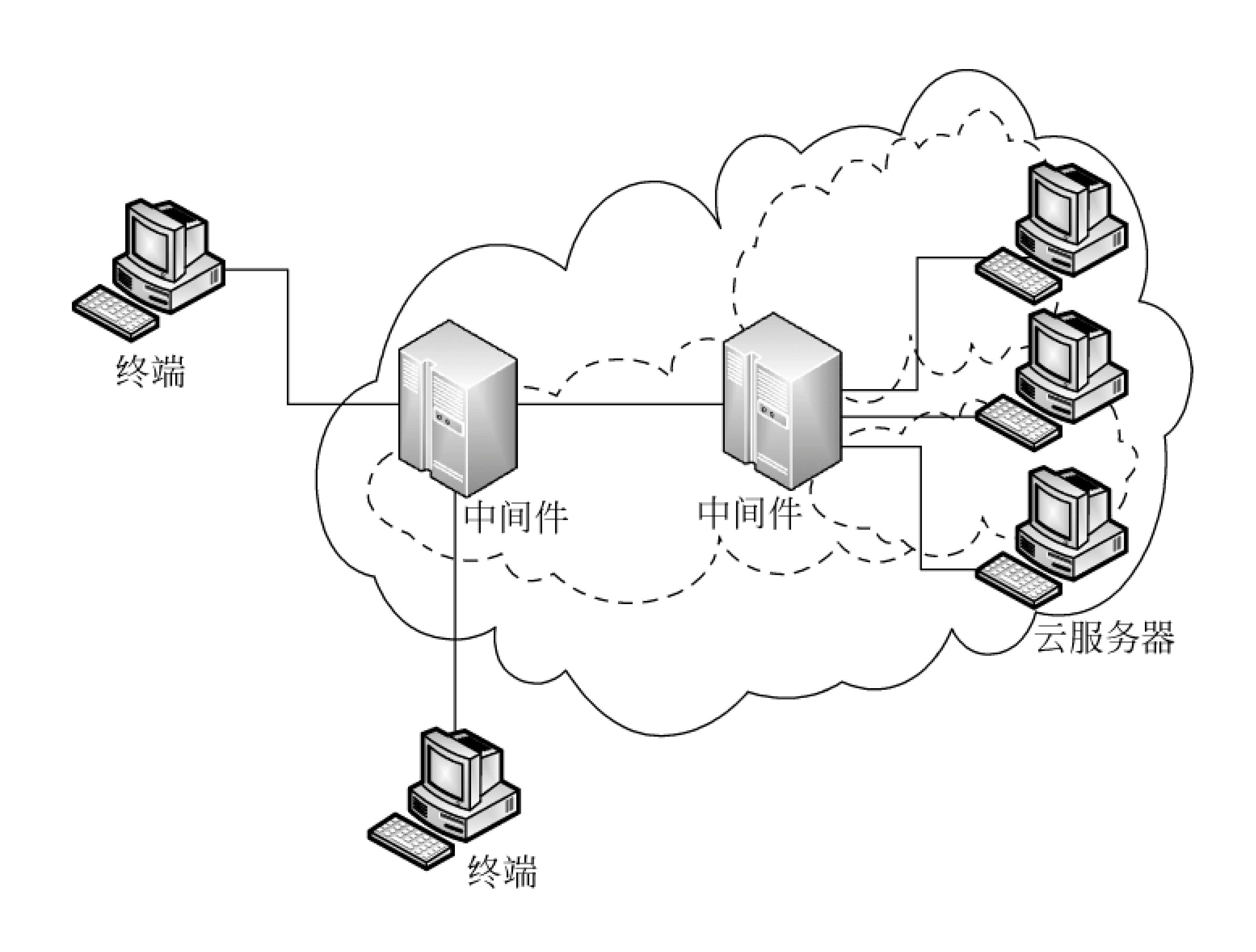


图 9-2　大数据系统的分布式云计算思路

9.2　大数据的分解模型

设社会网络大数据被划分成 C 个子模块。对第 i 个子模块，$i=1,\cdots,C$，u_i 为对第 i 个子模块的输入，x_i 为由其他子模块提供的中间输入，v_i 为对第 i 个子模块的控制变量，z_i 是子模块 i 对其他子模块的输出，y_i 为子模块的输出。以上各个向量，分别具有维数 m_{u_i}、m_{x_i}、m_{v_i}、m_{z_i}、m_{y_i}。这样的多级优化如图 9-3 和图 9-4 所示。

对于一个给定的总体输入向量 u，子模块可用下述向量方程描述：

$$\begin{cases} z_i = g_i(u_i, v_i, x_i) \\ y_i = h_i(u_i, v_i, x_i) \end{cases}$$

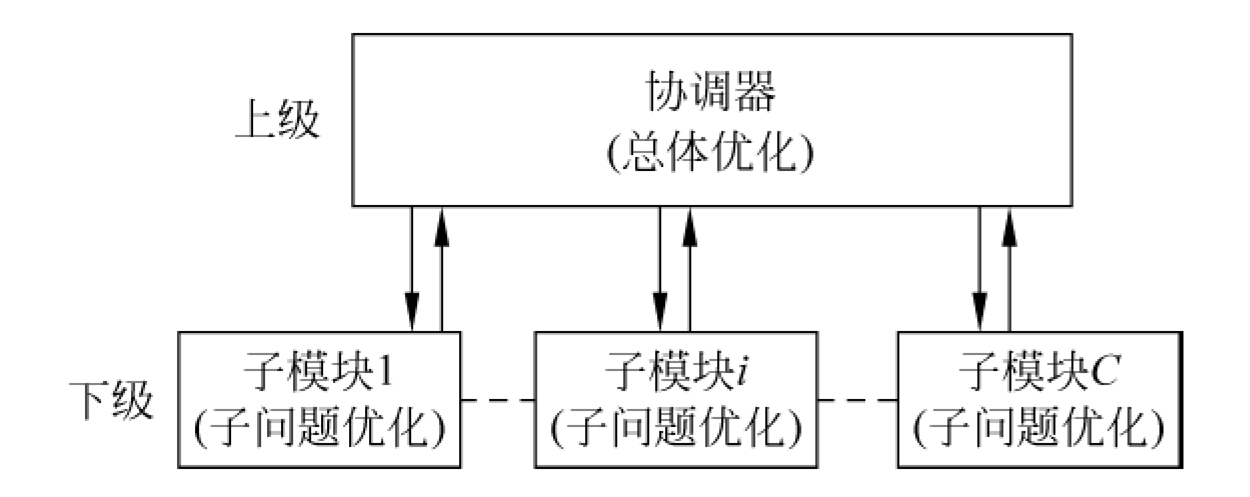


图 9-3　大数据的分解算法

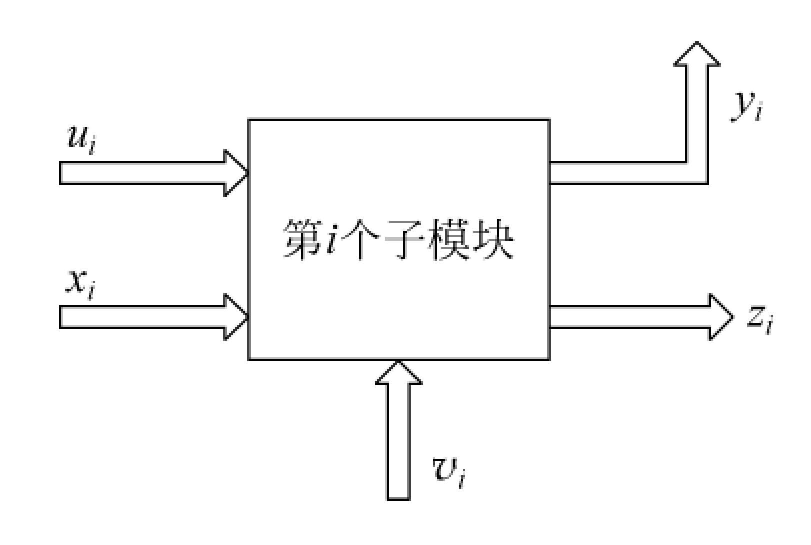


图 9-4　第 i 个子模块

子模块之间的关联如下：

$$x_i = \sum_{j=1}^{C} a_{ij} z_j, \quad i = 1,2,\cdots,C$$

其中，a_{ij} 为 $m_{x_i} \times m_{z_i}$ 矩阵，表达了子模块之间的耦合。

设大数据的目标函数是加性可分的，

$$\sum_{i=1}^{C} f_i(u_i, v_i, x_i)$$

写成拉格朗日函数

$$L = \sum_{i=1}^{C} f_i(u_i, v_i, x_i) + \sum_{i=1}^{C} \mu_i^T [g_i(u_i, v_i, x_i) - z_i] + \sum_{i=1}^{C} \rho_i^T \left(x_i - \sum_{j=1}^{C} a_{ij} z_j \right)$$

其中，μ_i 和 ρ_i 分别为 m_{z_i} 维和 m_{x_i} 维的拉格朗日乘子向量。设这些等式约束都是独立的，函数 f_i 和 $g_i(i=1,\cdots,C)$ 都是连续和一阶连续可微的，则最优解应满足下列必要条件：

$$
\begin{cases}
\dfrac{\partial L}{\partial x_i}=\dfrac{\partial f_i(u_i,v_i,x_i)}{\partial x_i}+\dfrac{\partial g_i(u_i,v_i,x_i)}{\partial x_i}\mu_i+\rho_i=0 \\[2ex]
\dfrac{\partial L}{\partial v_i}=\dfrac{\partial f_i(u_i,v_i,x_i)}{\partial v_i}+\left[\dfrac{\partial g_i(u_i,v_i,x_i)}{\partial v_i}\right]^T\mu_i=0 \\[2ex]
\dfrac{\partial L}{\partial z_i}=-\mu_i-\sum\limits_{j=1}^{C}a_{jj}\rho_j=0 \\[2ex]
\dfrac{\partial L}{\partial \mu_i}=g_i(u_i,v_i,x_i)-z_i=0 \\[2ex]
\dfrac{\partial L}{\partial \rho_i}=x_i-\sum\limits_{j=1}^{C}a_{ij}z_j=0
\end{cases}
$$

上述方程组形成了两级递阶结构的分解协调算法。在该算法中，上级和下级之间不断交换信息，下级子模块向上级送出反馈变量，上级协调器根据各子模块来的反馈变量，从全局优化的角度出发向下级给出协调变量，进行优化迭代，最后达到总体的最优点。

在上述变量中，可以采用不同的变量做协调作用，对应地形成了不同的分解协调算法，即目标协调法、模型协调法、混合协调法、三级分解协调及关联预估法。

9.3 网络舆情大数据面临的挑战

显而易见,网络舆情大数据面临着很多挑战。

1. 大数据增加了解决多学科集合问题的难度

威瑞森通信公司的迈克·博迪认为,当下我们面临的大数据挑战来自如何有效管理难以想象的海量数据以及如何将这些海量数据整合成我们所需要的有效信息,而不只是要酷似的玩弄技巧。事实上,在任何一个数据库中,非结构性数据(图片、声音和视频)所占的比例都越来越高,数据存储量从早先的 GB 已经发展到了 TB、PB 还有 EB,结构性数据在美国医疗保健大数据库中的比例已不足 10%,并且这一比例还在急速下降。大多数关联子数据库的语义格式并不兼容,因此大部分的数据分析仍然需要人工,这是实现以大数据为基础的“集成问题解决方案”的难点所在。

2. 海量数据增加了有效数据的使用难度

当下的 Web 3.0 时代是“基于数据的网络”时代,互联网已经成为一个超大的关系型数据库,其特征表现为:个性为主;强调用户体验;良好的模块制定功能;数据整合能力强(周珍妮,陈碧荣,2008)。据统计,现有数据网络含有 310 亿个 RDF 三元组,其中 4000 多万个 RDF 连接的三元组将不同数据元之间的数据串接起来。这些数据中政府数据占 41.9%、地理型数据占 19.4%、出版和媒体类数据占 14.8%以及生命科学数据占 9.7%。

原始的大数据呈现出一片混乱的状态。从事数据工作的人普遍认为80%的精力都用在了数据清理上，正如彼特·沃登在其著作《大数据词典》中所言："我可能花更多的时间整理那些杂乱的源数据，而不是直接就开始分析数据。"

数据网在以下三个方面为数据整合和大数据处理增加了难度。

（1）通用和专有词汇的使用，例如"人"、"商品"这一类常见的表达，关联数据资源是可以借用的，但其他常见表达里没有的词汇关联数据资源需要自定义。借用更多广泛运用的常见表达词汇，可提高不同数据资源的通用性。

（2）不同格式数据对同一对象描述的认定。不同计算机语言之间对同一对象的描述可能是不同的，应用程序如果能辨识同一对象在不同语言中的表达将有助于数据集合和数据清理。

（3）由于媒介平台的开放性，自媒体时代人人都在发布资讯，大部分的互联网数据都是垃圾数据，因此科学评估数据质量和确定有价值的数据子集也是一大挑战。

3. 大数据平台需要有可以处理不同种类数据的数据整合技术

Openlink公司的首席软件设计师奥瑞·俄凌指出目前人们已经意识到了智能数据处理的前景，但现实使用情况几乎还是空白。类似现在运用的OWL语言可能是数据融合的处理方式之一，但不会是未来的方向。目前的关联数据和RDF在数据整合技术中占有一席之地，它们的国际通用性强，且为无预定数据模式。

资源描述框架（RDF），是一个用于表达关于万维网上的资源信息的语言。它专门用于表达关于 Web 资源的元数据，比如 Web 页面的标题、作者、修改时间以及 Web 文档的版权和许可信息，某个被共享资源的可用计划表等。RDF 使用 XML 语法和 RDF Schema 来将元数据描述成为数据模型。数据对资源的描述是与领域和应用相关的，比如对一本书的描述和对一个 Web 站点的描述是不一样的，即对不同资源的描述需要采用不同的词汇表。一个 RDF 文件包含多个资源描述，而一个资源描述是由多个语句构成，一个语句是由资源、属性类型、属性值构成的三元组，表示资源具有的一个属性。RDF 规范并不定义描述资源所用的词汇表，而是定义了一些规则，这些规则是各领域和应用定义用于描述资源的词汇表时所必须遵循的。通过 RDF，人们可以使用自己的词汇表描述任何资源，由于使用的是结构化的 XML 数据，搜索引擎可以理解元数据的精确含义，使得搜索变得更为智能和准确。

9.4 网络舆情大数据发展方向的展望

1. 大数据的应用领域

1）经济管理

与大数据相关的技术为经济管理创造价值提供了重大的新机遇。零售部门不仅记录下每一笔交易和操作，还记录着新出现的数据源比如 RIFD 芯片，可追踪货物、在线消费者的行为和感情表

现，这使得数据量的增长势不可挡。麦肯锡全球研究所的大数据报告介绍了大数据技术在零售业、制造业中的应用及其对整条行业链的影响(James et al.，2011)。

(1) 零售业。

事实上，零售业通过使用信息技术的影响力获利的做法已经有几十年的历史。比如，在美国，零售终端的交易数据主要从条形码中获得。条形码在20世纪70年代首次出现，20世纪90年代之后，许多大型零售商都开始使用门市层级和供应链的数据来优化配送和物流，加快货物规划和管理，升级店铺运营。比如沃尔玛研发的“Retail Link”可以让供应商大致浏览其门店，了解什么货品需要重新进货而不是被动等待订单。这种“厂商管理存货”的方法是一个革新性的概念，在20世纪80年代开始使用。沃尔玛尝试不间断的管理创新方法，直接和间接地促使了整个日用百货行业在20世纪90年代的生产力加速提升，如仓储式格式，每日最低价，提升竞争强度，鼓励最优的管理和技术的扩散。自此之后，其他零售商开始模仿沃尔玛，以保持竞争力的同时，整个行业的生产效率随之全面提升。

今天，领跑者们正在挖掘消费者数据，为从管理供应链到推销和定价等一系列问题提供决策参考。消费者数据日益颗粒状，这些数据来自多种销售渠道、商品目录、商店、在线互动。随着整个行业对大数据的认识的加深，零售商将大数据工具应用到运行和供应链，可持续降低费用，不断创造新的竞争优势和策略，并获得

更大的效益。

（2）制造业。

制造行业是大数据早期的和重度的使用者，在电脑诞生之日就开始使用信息技术和自动化技术来设计、制造和配送产品，目的是提高产品质量和性能。在 20 世纪 90 年代，制造业公司获得了惊人的年度生产能力的增长，因为运行的改进提升了制造过程的效率，也提高了制造产品的质量。制造商还优化了全球运行和管理，将产品外包给成本更加低廉的地区。相对于绝大多数行业，制造业相对已是非常高效，但是大数据仍然能够提供另一波重大的制造业升级。

2）社会管理

当前我国社会处于经济快速发展时期，同时也凸显各种社会矛盾。政府需要收集巨量数据与数百万公民打交道，绩效的表现也是参差不齐。面对大数据这一潜在的宝库，却很少有管理者主动发觉所拥有的信息，而政府往往将数据保存在各自为政的部门中。政府部门是否可以通过大数据的应用提升自己的生产力和工作效能呢？麦肯锡研究了欧盟国家的政府部门行政管理，发现大数据的应用工具可以为社会管理提供有效的策略和技巧，以提升生产力、提高效率及影响力。欧盟政府部门可能会减少 15％～20％的行政开支，创造 1500～3000 亿欧元的新价值，大数据还可以在未来 10 年中将年度增长率最高提高 0.5％。

麦肯锡全球研究所的大数据研究报告显示，欧盟国家对大数

据工具的应用可以从四个主要方面推动社会管理水平。

（1）实现信息透明化。

若政府部门大数据库的数据更加易得，外部利益相关者（如公民和企业）以及内部利益相关者（如政府雇员和政府机构）都能够提高自身的工作效率。目前越来越多的不同层级的政府部门开始引入“开放数据”原则，允许公众获得原始政府数据。这样的努力开启了海量的数据创新，人们将多种来源的数据结合起来以创造类似“网络城市”新闻，记录在某个特定城市发生的事件。

（2）发现需求、展现差异和提高绩效。

大数据的重要贡献之一是它可以发现不同政府机构在行使相似职能时呈现出的巨大绩效差异，这个信息对在机构内提高各部门的执行能力提供了重要的机遇。对于政府部门这类外部竞争压力较弱的组织，凸显不同机构、部门工作绩效可以带来内部竞争、提高效率。

（3）人口细分和制定政策。

麦肯锡的研究报告发现，根据个体和人群将公共服务进行细分与制定能够提高效率、效果和公民满意度。同样，政府的税收部门可以使用大数据对个人和企业纳税人进行分割，比如，可以将纳税人按照地理、守信记录、违约风险、收入水平等特征分类。有效的分割可以将潜在征缴和实际征缴之间的差距缩小10％，同时更加精准的互动还可以将用户满意度提升15％。

（4）使用自动计算代替或辅助人为决策。

大数据的更为复杂、更为高级的应用是使用自动算法来分析大数据库，从而帮助决策者判断。运算法则能够从多种源头抓取大量数据，识别出不一致、错误和虚假信息。

3）医疗健康

目前，医疗系统在提高运行绩效和采用科技辅助过程方面落后于其他许多部门。改革现有的医疗制度，削减医疗成本的增长率，同时还要维持现有的优势，这是全球各个国家社会和经济共同面临的关键问题。

鉴于此，使用大数据作为工具，将会为生产出更有效、更加经济的医疗政策、更高的产品和服务提供新的商业模式。根据麦肯锡的预测，在医疗领域具备所需的 IT 和数据库投资、分析能力、隐私保护以及适当的经济激励机制的情况下，大数据的使用将在 10 年内让美国的医疗市场获得每年 3000 亿美元的新价值，其中 2/3 以全国医疗开支的削减形式出现。

美国以及欧盟在临床、支付与定价、研究与开发、公共健康等领域中已经涌现出多种大数据技术，能够利用医疗部门中已有或可能获得的海量电子信息提高医疗系统的效率和效果。

（1）临床。

在临床方面，如果采取结果导向的疗效比较研究，可以通过分析详尽的患者和治疗结果信息，比较不同方案的效率，从而决定针对特定患者的最佳治疗方案。推行医疗比较系统，很有可能减少

过度医疗和处理不足的发生率，这两者都会致使患者状况恶化以及产生更高昂的长期治疗费用。

临床决策支持系统可以提高手术及医嘱录入系统的效率和质量。通过使用医嘱录入系统，医疗服务提供机构能够减少不良反应，降低错误治疗和民事诉讼的比率，特别是降低医疗事故的发生率。

（2）支付与定价。

自动化系统可以识别欺诈，并核实支付者补贴申请的一致性和准确性。同时，基于真实患者治疗效果数据，使用卫生经济学和效果研究的定价方案，可以实现公平的经济补偿。

（3）研究与开发。

在制药的子领域，大数据工具可以提高研发的生产力。它们可以共同创造高于1000亿美元的价值，其中1/4的形式为更低的国家医疗保险费用。具体的大数据工具包括预测模型、统计工具和算法式改善临床试验设计、分析临床试验数据、个性化药物以及分析疾病模式等方法。

（4）公共健康。

大数据的应用能够改善公共健康监视和反馈。通过使用全国范围的患者和治疗数据库，负责公共健康的政府部门能够保证快速、协调地发现传染性疾病，全面监视疾病爆发，制订完整的疾病监测和反应计划。

2. 网络舆情大数据发展方向的展望

1）探索基于云计算和大数据结合的最优方案

半个世纪信息技术的发展，我们主要解决的是云计算中“结构性”数据的存储、处理与应用。“结构性”数据的最主要的特征是逻辑性强，每个“因”都有“果”。然而现实社会中大量数据事实上没有显现性的因果关系，如一个时刻的交通堵塞、天气状态、人的状态（心理与物理）等，它的特征是随时、海量与弹性，如一个突变天气分析包含会有几百个 PB 数据。而一个社会事件如乔布斯去世瞬间所产生在互联网上的数据（微博、纪念、文章、视频等）也是突然爆发出来。大数据时代就是这样一个以 PB 为单位的结构与非结构数据信息组成的包含社交网络、电子商务与移动通信等体系的互联网时代。

大数据的一个特点就是海量。海量的数据需要足够存储来容纳它，快速、低廉价格、绿色的数据中心部署成为关键。近几年，谷歌、Facebook、Rackspace 等公司都在纷纷建设新一代的数据中心，大部分都采用更高效、节能、定制化的云服务器，用于大数据存储、挖掘和云计算业务。海量数据，就是大数据和云计算的交集。

云计算中的大数据有几个核心要素，如数据在云端的集合与分享、个人数据的无缝连接（随时、随地、同步）以及数据的跟踪分析和挖掘。一方面由于云计算拥有可以弹性扩展以及相对便宜的存储空间和计算资源的特点，中小企业、机构也可以通过云计算完成大数据的分析；另一方面云计算 IT 资源庞大、分布较为广泛，

是异构系统较多的组织及时准确处理数据的有力方式。云计算与大数据的关系是两个方向，云计算可以承载大数据，大数据也是可以通过云计算架构和模型来提供解决方案。也就是说，大数据在管理和应用的方向上，可以通过云计算的资源共享、高可扩展性、服务特性来搭建和运营。

然而，随着非结构化数据比例的增加，传统为结构化数据存储而设计的存储系统，已经无法应付云平台系统庞大的数据存储需求。云存储服务中，这些数据保存成本高昂，它们的移动(存入及取回)也存在很大的困难。如何解决不同的云存储环境中结构化、非结构化数据的处理和应用问题，怎样为这类问题提出优化的、可行的技术与管理方案，已经成为大数据决策研究的重点之一。

2）探寻大数据内部关联和挖掘的新方法

如果说大数据与云计算的交集是外部交集，那么大数据内部的关联、挖掘，则是大数据的大内涵，这个调整远远超过云计算的应用难度，数据与数据的复杂关系，比如跨应用系统的结构化数据与非结构化数据的关联，海量数据的存储以及数据在人之间的分享，数据(结构化与非结构化)与业务和决策间的关联等。

解决这个问题的一个思路是 EMC 提出的信息生命周期思想，信息生命周期管理作为一种信息管理模型，认为信息有一个从产生、保护、读取、更改、迁移、存档、回收的周期、再次激活以及退出的生命周期，对信息进行贯穿其整个生命的管理需要相应的

策略和技术实现手段。信息生命周期管理的目的在于帮助企业在信息生命周期的各个阶段以最低的成本获得最大的价值。但是，在很多业界同行看来，这个思想在云计算和大数据时代需要真正的升华，而这个方向就是智能，不是单一和局部的，而是统一的智能。

作为国内崛起的新兴代表，爱数则提出了智能数据管理解决方案，也是基于统一智能和信息生命周期思想的框架。无疑，这是一种新的思路，这种统一的框架采用云计算体系结构，主要技术突破在资源池化和法规管理遵从，从数据生成阶段就将大数据纳入到生命周期管理中，通过统一的智能策略，既提供了很好的运维和保护，也能在使用和挖掘阶段与业务应用结合起来，提供统一的数据信息平台。大数据内部的关联和挖掘问题是大数据未来发展道路上的一个挑战，若能击破这个问题，这将会是我们在大数据的研究与应用上质的飞跃。

3）探索大数据复杂性、不确定性特征描述的刻画方法

数据内部关系的复杂性以及不确定性给大数据特征的描述带来挑战。对于一个复杂数据，如文本，视频，图像，或者生物实验数据，人们需要从不同的角度去诠释这样的数据。例如，在网页数据中既有关于内容的文本属性，也有指向这个网页的链接属性。同时大数据不确定性较强的特点，也给数据特征的描述与刻画增加了难度。

数据分析家们已经有了这样的共识，那就是以前的单维聚类

方法不再适合大数据的多样性特征。有学者提出多维聚类分析方法，通过对单维聚类问题的扩展，为复杂数据提供了一种新的探索性分析的方式。多维聚类也只是在大数据划分上提出的思路，对于大数据有效、明确清晰的挖掘和诠释方法还需要人们继续的探索。

4）研究大数据时代对管理决策的影响

随着世界开始迈向大数据时代，社会也将经历类似的地壳运动。在改变我们许多基本的生活和思考方式的同时，大数据早已在推动我们去重新考虑最基本的准则，包括怎样进行管理与决策。在一个可能性和相关性占主导地位的世界里，专业性变得不那么重要了。专家行业不会消失，但他们必须与数据表达的信息进行博弈。大数据的背景下，直觉的判断往往被迫让位于精准的数据分析，这将迫使人们调整在管理、决策、人力资源和教育方面的传统理念。

从 2004—2012 年间担任美国统计局和商务部的高级顾问胡善庆，指出一些国家已开始建造有关就业、教育和公共卫生的公众纵向数据计划。这些计划虽在不同的发展阶段并且仍然具有很强的挑战性，但它们提供了在大数据时代建造和维持广泛、详细动态统计系统是可行的这一令人鼓舞的消息。大数据不只是反映现代科技进步对改善统计计算的需求，它是向传统统计专业的一场挑战，并要鼓舞创新思维和发展的一场大革命。

9.5 本章小结

大数据不仅增加了解决多学科集合问题的难度，加大了有效数据的使用难度，而且大大提高了不同种类异质数据的整合难度。本章提出了问题的环境和解决问题的思路及框架，研究了大数据的分解模型，展望了网络舆情大数据发展方向。

思考题

1. 熟悉解决问题的环境和解决问题的思路及框架。
2. 掌握大数据分解模型的基本思路。

第10章

企业社会网络舆情管理方法

本章学习目标

- 了解社会网络下 C2B 营销的实现及其对企业业绩的影响。
- 熟悉企业的社会网络个性化信息推荐等企业舆情管理方法。

本章在前面基础模型和衍生模型的基础上，通过对模型的多种组合，得出针对企业在社会网络大数据环境下的舆情具体管理方法与策略，力图解决目前企业在新型社会网络大数据环境下遇到的舆情管理方法的几个典型问题。

10.1 社会网络下 C2B 营销的实现及其对企业业绩的影响

B2C(business to customer)是由企业到消费者的传统营销模式，而 C2B 正好与之相反，即 C2B 是从消费者到企业的商务行为。

在传统的互联网下，对众多的互不相识的用户，要完成 C2B 商业行为，是难以想象的。只有在 Web 2.0 环境下的社会网络中，用户个体之间以及用户和企业之间的交互才能够得以实现，也正是这种环境才成就了 C2B 的企业盈利模式。

显然，在传统环境下，如果一个用户向企业定制一件特殊的、但企业尚未生产的“个性化”产品时，企业需要为此进行专门的设计和生产，从而显著增加了制造单件产品的生产成本，导致了销售价格昂贵。而在 C2B 模式下，通过汇聚具有相似或相同需求的消费者，形成一个特殊群体，一大批用户一起定制某个特殊产品，就会使得单个产品的生产成本下降，同时消费者经过集体议价，也可以达到消费者购买数量越多，价格相对越低的目的(对已有产品的 C2B 营销模式称为传统网络团购，因此传统网络团购只是 C2B 营销模式的一种)。

因此企业需要及时发现和了解在社会网络中，由于对某件未生产的产品感兴趣而逐渐聚集的团体。C2B 企业可以使用模型 Ⅰ-1 对用户情感进行监控，还可以加入讨论，并进行企业舆情管理的优化。

以消费者为中心、消费者参与设计与生产、由消费者主导等属于是 C2B 营销模式的主要特征，其本质特征是先有消费者需求，后有企业设计并生产。C2B 的经济关系被视为是一种逆向的商业模式，能够通往大众的双向交流人际网络使得这种类型的商业关系变得可能。在 C2B 的企业盈利模式下，企业生产和运作的上下游

关系倒了过来，消费者由下游变成了上游，成为企业的“龙头”和“风向标”。

这种基于社会网络的新的C2B盈利模式，对购买者、供应商、潜在进入者、替代品、行业内竞争等都产生了巨大的影响，供应商、渠道建设、信息流、资金流也有着自身的新特点，相应的营销管理、客户关系管理、生产管理、信息管理、财务管理都需要在广义企业舆情管理的驱动下随之创新。

10.2　企业的社会网络个性化信息推荐

社会网络个性化信息推荐可以分为用户之间的信息推荐、以及企业对用户的信息推荐。管理者可以和发现的权威用户建立友好关系，经常转发或评述与企业直接和间接相关的信息；而对于企业的负面信息，力争使关键用户的邻居少进行或不进行扩散，使得企业舆情对该企业有利，从而达到“四两拨千斤”的效果。

社会网络中潜藏着用户之间的信任关系，用户之间的连接关系的强弱反映了用户之间信任关系的强弱，从某种意义上也说明推荐中潜藏着用户之间信息传播的路径，这个路径要比其他的行为路径关系更快捷、更有目的性。因此，挖掘社会网络中的个性化推荐行为之间的联系，必然能寻找出用户及用户群体之间的深层次的依赖关系。

除了有效的监测和捕获舆情外，对于掌握舆情的发展过程以

及对舆情的引导和堵截也很重要(Tian 和 Liu,2014),这需要企业掌握社会网络的结构上的特性。在社会网络中,企业也可以通过挖掘用户之间的信息推荐等行为,来挖掘出用户是通过怎样的路径、哪些用户、哪些方式来传播信息,这对于企业社会网络舆情的管理是非常有必要的。

10.3 社会网络大数据环境下企业的开放式信用管理

在社会网络大数据环境下,企业的信用开放式管理问题不同于传统环境下的企业信用管理。社会网络大数据环境下的企业关于用户信用资源的获取渠道更加广泛,数据量也更加庞大。同时,企业与用户之间的交流也更加密切,因此需要建立新的信用管理与评估机制,对用户在社会网络大数据环境下的开放信用信息进行抽取和评估。

在社会网络大数据环境下,通过建立基于企业社会网络舆情的社会网络信誉平台,用户的信用信息主要来源于用户在社交网络中相应行为的开放式获取,包括用户的关注主题倾向与用户的动态交互评估。

在研究中,我们可以利用用户的社区发现模型(模型Ⅱ-1)和社会网络舆情计算中的用户影响行为模型(模型Ⅱ-3),来对用户在社会网络大数据环境下的信用问题进行分析,得出企业在社会

网络大数据环境下对用户信用信息的最佳管理方案，提升企业传统信用管理的效率，降低管理成本。

10.4　社会网络大数据环境下针对在线舆情服务挽回管理措施对企业绩效影响评估

实践中主要有两个做法：一是及时发现用户抱怨产品或服务信息，通过企业客服的在线交流，满足用户的退换货要求，并在网络上及时把处理结果公布，从而提高企业信誉度和可信度，进一步提高企业利润。二是发现负面舆情，例如高管被带走，及时发布企业正面的积极响应信息，将企业公众形象减少到最低，企业经济损失减少到最小。

社会网络大数据环境下的舆情管理措施对企业绩效的影响，需要通过对企业实施舆情管理前后的绩效评估变化来进行分析。不同于传统的绩效评估方法，基于社会网络环境下企业的绩效评估还应该包含企业网络声誉（舆情）以及企业的网络价值评估等，这些方面的绩效评估策略可以利用我们已有的用户动态行为和影响行为模型（模型Ⅱ），结合企业原有的绩效评估方法进行定量的分析。同时，为了能够有效地探讨优化后的在线舆情管理方法对企业绩效的影响，我们还可以在企业绩效的特定考核上设置舆情绩效的监控点，对比不同的组合优化方法导致的企业绩效的定量变化，以找到在线舆情管理与企业绩效的交互机制。那么，不同类

型的企业就可以得出不同的优化算法(参数与模式匹配后)上,得出更加符合自身的最佳舆情管理策略的优化组合,实现企业舆情管理的效益最大化。

当舆情发生后,社会网络中的少量的用户将信息传递给更多用户,造成舆情事件消息的进一步扩散和关注度迅速升温。根据模型Ⅰ,我们可以对企业舆情的传播内容进行分析,包括分析传播内容的情感倾向性。利用模型Ⅱ,我们还可以对这些少量信息的传播范围和路径进行分析,找出传播最快的社区及关键路径。采取公关措施,阻止坏的消息传播,推动好消息的传播。在关键路径上、权威用户处播放有利于企业的消息,完成企业的舆情处置。

10.5 社会网络大数据环境下企业舆情管理方法及其对在线舆情的影响

企业舆情管理既是企业在社会网络大数据环境下,通过技术手段针对开放式网络舆情按照企业目标进行控制的过程,也是企业内部管理过程在网络舆情的作用下自主改变和演变的过程。因此,企业舆情管理和在线舆情间是一个相互影响相互改变的关系。

在企业对舆情管理的方法方面,可以按照舆情的事态特征分为日常舆情管理和突发舆情管理。基于模型Ⅰ-1,通过对开放网络的实时信息抽取,实现对于舆情的实时发现。对于日常舆情,结

合基于社会网络拓展的施拉姆理论的舆情处置管理方法，根据模型Ⅰ，对日常舆情采用引导管理模式，进行内容提取和舆情引导，并按照内容将其反馈并进入企业内部管理流程，通过管理，使得日常舆情能够对企业经营绩效产生正面影响。对于突发舆情，结合基于社会网络拓展的生命周期理论的舆情处置管理方法，根据模型Ⅱ找出社会网络中的社区、传播的关键路径、权威用户等，确定舆情传播方向和爆发、影响范围，启动应急响应机制，使企业对突发舆情采用的措施更具有针对性，更能优化步骤。然后，结合基于社会网络拓展的价值累加理论，实施关键用户控制，在关键信息披露、关键时点管控、关键群体影响等方面，使得企业能进行更有效的舆情调节和控制，以实现舆情对企业经营和业绩的最小冲击，实现企业管理和绩效与在线舆情的良性互动。

10.6　本章小结

本章讨论了社会网络下C2B营销的实现及其对企业业绩的影响、企业的社会网络个性化信息推荐、社会网络大数据环境下企业的开放式信用管理、社会网络大数据环境下针对在线舆情服务挽回管理措施对企业绩效影响评估及社会网络大数据环境下企业舆情管理方法及其对在线舆情的影响等企业舆情管理方法。

思考题

1. 举例说明社会网络下 C2B 营销的实现及其对企业业绩影响的应用。

2. 举例说明企业的社会网络个性化信息推荐的应用。

第11章

展　　望

本章学习目标

- 了解企业社会网络舆情给企业管理的挑战。
- 了解企业在社会网络大数据环境下的舆情管理应对策略。

11.1　企业社会网络舆情给企业管理的挑战

随着 Web 2.0 技术的广泛发展与应用，网络用户日益成为网络内容的重要创造者之一，同时深刻地改变了企业经营、管理所面临的环境（Andriole，2010）。利用在线平台提供的信息发布功能，网络用户可以轻松发布与企业相关的信息，一次表达自身的信念、态度、意见和情绪，从而形成狭义的企业舆情。比如通过微博客、

论坛，用户可以发表关于企业动态的博文、帖子，表达自身的看法与见解；通过电子商务平台，用户可以发表针对企业产品的在线评论，反馈产品的使用体验与改进意见。同时，由于网络应用的社会性特征，企业相关的信息会被推送至发布者的网络好友和关注者，并通过转发、跟帖和分享等操作被传播至更广的范围，从而给企业信誉，甚至经营管理带来巨大的影响。在新型社会网络大数据环境下，企业无法忽视这类网络社会舆情给企业自身营销管理、客户关系管理、生产管理、信息管理以及财务管理等诸多方面所带来的重要影响，同时企业管理者更不可能躲在“世外桃源”的环境里，置身事外。相反地，企业的管理者必须正视舆情，从而学会管理、控制，甚至引导舆情，并尽量降低负面舆情对企业的影响。因此，企业需要在理清社会网络大数据环境下企业舆情管理特征的基础上，对企业的原有管理模式进行适应性的改变。

第一，企业在线网络舆情的自媒体、社会性特征给企业的品牌资产（brand equity）管理（Aaker，2009）带来了前所未有的机遇，同时也提出了更加严峻的挑战。相比 Web 1.0 时代网络用户的浏览者身份定位，如今的网络用户更多地扮演了信息生产者、传播者的角色。而且相比于传统的广告等企业推送的信息，网络用户自主创造、加工、传播的企业舆情更能得到其他网络用户（好友）的信赖，从而产生更广、更强的、更有价值的影响（Archak 等，2011；Chen 和 Xie，2008）。企业在线网络舆情传播的这些特征为企业扩大品牌知名度（brand awareness）（Keller，2003）、提升品牌认知度

(perceived quality)提供了便捷、有效的渠道。另一方面,网络也使得企业的负面消息、新闻能够得到更快、更广地传播,形成不利于企业的网络舆情。而且相比于正面的企业网络舆情,负面企业网络舆情更容易受到网络用户的关注,并更能影响网络用户对企业的认知(Bambauer-Sachse 和 Mangold, 2011; Zhu 和 Zhang, 2010)。因此,当出现负面的企业网络舆情时,如何及时、准确地采取恰当措施阻止或延缓负面舆情的发展、传播,减小负面舆情对企业品牌资产的影响,成为企业危机公关(crisis communication)(Fearn-Banks,2010)中面临的重要问题之一。

第二,社会网络的出现使得企业决策者可以在企业市场细分等方面实施新的战略。社会网络上各种具有某种兴趣或偏好的团体更易于出现和被企业所掌握,此时企业的营销管理就需要"与时俱进"。与传统的客户分类不同,在现今的社会网络中,用户的分类很难进行准确的界定,市场呈现出一种碎片化和多元化的状态。套用"长尾理论"就是:长尾部分商品将带来巨大的利润。换句话说,这些在原有的市场分类模式中,处于碎片化或称"散户"的用户可能会给企业带来大量的销售利润,而不再仅仅是传统意义上的大客户销售(或管理)的问题。这正需要进一步研究,这也是近年来一个流行于互联网企业口号"得草根者得天下"的根本体现。这时,如果企业能够利用微博、微信、即时通信软件等社会网络平台对用户的社会网络活动及其他偏好信息进行收集、处理和智能分析,以此精准了解用户的产品喜好,实时掌握用户的需求动向,深

度过滤用户的特点和购买意图等。

第三,社会网络大数据环境下的企业舆情也给企业在客户获取与保留(客户关系管理)(customer acquisition and retention)方面提出了新的要求。社会网络商务环境下的企业客户既是传统意义上的消费者,也有作为企业产品创作者的成分。不同于传统商务模式中零散的买家结构,在社会网络背景下,买家更容易形成利益团体,从而在一定程度上降低信息的不对称性,提升议价能力等等。因此,企业可以利用社会网络的用户社区与客户及潜在客户建立更为深入的联系,这与社交网络中的虚拟社区是一致的。企业通过建立产品社区、兴趣社区、互助社区等,与用户建立直接、高效和实时的联系渠道,为用户提供交流、评论和反馈的平台,以此进一步促进客户关系的提升与管理优化。

第四,在价值网络上,社会网络大数据环境下的企业舆情管理功能包括提升企业产品和服务在客户中的认知水平、帮助客户评估企业的价值主张、协助客户购买特定产品和服务以及提供及时有效的售后客户支持等。通过文本挖掘、社区发现等智能数据挖掘分析方法可以找出社会网络大数据环境下用户对企业产品和服务的讨论,然后由专门的人员进行回应或引导,从而提升产品或服务在用户内心的认知。例如,在微博社会网络中,普通的微博用户通过关注好友的微博了解某些企业的产品或服务,以此来评价企业价值主张。企业也可以通过合理的微博营销手段(比如微博大V的推荐等)向客户传递自己企业的价值主张,对自身企业进行宣

传和营销，以此改善与用户联系的渠道通路。

第五，在线企业舆情也给企业的供应商选择提供了更多的可能选择以及更加全面的信息。通过分析上下游相关企业的在线舆情，企业可以更加详细地了解、评估相应企业的生产能力、产品品质和业内口碑，从而更加合适地选择合作伙伴。企业舆情为企业选择合作伙伴提供了更多的选择，以及更加丰富的参考信息（Carter 和 Rogers，2008）。相应的，供应链上下游企业也会通过企业舆情来判断企业的生产经营状况和信用情况。因此，整个供应链上下游的企业均可以通过在线舆情了解彼此的信息，并基于企业舆情做出供应链合作的决策。这就使得企业在做出违约等破坏供应链合作的行为时，必须考虑该行为对企业舆情带来的影响，以及进一步对企业今后的合作带来的影响。企业舆情给供应链上下游企业间的合作提供了一种无形的约束机制（Dellarocas，2003）。

综上所述，社会网络大数据环境中的在线舆情改变了企业的经营、管理环境，使得企业在面对市场、消费者、上下游企业和内部管理时均需要考虑企业舆情带来的变革。要充分利用企业舆情带来的机遇并应对相应的挑战，首先需要拥有相应的技术方法能够从海量的富媒体（rich media）数据中识别出与企业相关的信息，并分析企业舆情中网络用户所讨论的话题内容以及情感倾向，从而刻画出企业舆情的状态及动态变化走势。基于企业舆情状态与走势的刻画，企业便可进一步分析企业舆情状态与企业经营管理绩效、股票市场表现等的关系，从而判断企业舆情走势对企业表现的

影响。同时，基于企业舆情走势情况，企业可以得到企业舆情传播、发展的规律，并采用相应的手段、方法管理企业舆情的传播与发展，以保证企业经营管理目标的实现。

11.2 企业社会网络大数据舆情管理的应对策略

社会网络大数据环境下的企业在线舆情也为企业的产品设计与开发提供了宝贵的市场反馈信息(market feedback)。社会网络大数据环境下的企业舆情中包含了消费者大量关于企业产品和服务的使用体验以及相应的意见与建议，反映了企业产品的优势与缺陷(Zhang 等，2012)，为企业了解其产品的市场接受程度(market acceptance)提供了便捷的渠道(Chen 等，2011)。相比于传统的电话回访、问卷调查等方式，在线企业舆情分析可以帮助企业以更低的成本更加真实、准确地获取消费者关于企业产品和服务的意见反馈，从而有针对性地改进产品设计和服务。另一方面，企业还可以通过分析相关企业的在线舆情，了解相关产品的市场接受情况，以及消费者对于本企业产品和相关企业产品的比较评价(Xu 等，2012)。通过比较企业间的在线舆情，可以帮助企业识别竞争对手、判断竞争程度(Clark，1999；Ketchen 等，2004)，并根据消费者意见采纳竞争对手产品的合理设计。

例如，在微博中，有针对某个话题进行讨论而临时建立的话题平台(通常是以“#话题#”的形式存在)，这种临时的话题平台虽

然不如专业社区功能完善，但也将对一些企业或产品有相同兴趣的用户临时的集中起来，客户可以在这个话题下进行讨论，从而获得信任或权威的其他用户对于商品或服务的相关意见，以此参与到产品的营销和销售互动中来。而企业的管理者则需要敏锐的发掘、管理甚至制造这样话题平台，从而可以与原有的固定社区一样帮助、管理和促进客户关系的提升。

目前，很多企业都建立了微博公众平台、微信公众平台、企业QQ 等社会网络媒体官方窗口，并由专门的人员进行管理。例如在平台上发表和“收听”消息，通过和相关人员交流、参加社区讨论等。用户在产品使用中的问题等都可以通过这些平台与企业沟通，企业人员可以在平台上对用户的评价和留言进行快速的回复处理。这样就扩展了原有的只能通过电话和邮件与售后或客服人员联系的方式，使厂家可以与客户建立更加紧密迅捷的关系。

在新兴社会网络大数据环境中，网络用户倾向于在具有相同兴趣、爱好的用户间接收并传递内容，在分享企业及其产品信息的同时传递产品偏好。相比于传统的媒体广告等方式，利用网络用户间自发传递的企业舆情来进行客户获取，能够使企业更加准确的定位目标客户，获得更高、更长远的客户资产和企业收益(Villanueva 等，2008)。

企业舆情在客户获取方面的突出表现，使得企业必须关注新环境下基于企业舆情的客户获取策略。一方面，企业舆情中反应了网络用户关于企业产品或服务的意见情况，为企业了解客户的

满意程度提供一种新的途径。而且,网络用户所处社会网络中相邻用户关于企业产品的态度和行为也有助于企业准确预测该用户是否会放弃本企业产品(Nitzan 和 Libai,2011)。另一方面,企业需要在社会网络大数据环境中基于企业舆情分析来预测客户流失(customer churn)情况,并有针对性地设计客户挽回措施。

鉴于社会网络大数据环境中的企业舆情给企业经营、管理带来的重大机遇和挑战,以及问题本身的理论意义和价值,我们可以针对现有企业舆情研究中的不足和缺失开展相应的研究,探索新兴社会网络大数据环境下在线舆情对企业绩效的影响机制及企业管理网络舆情的理论与方法。

11.3　本章小结

本章讨论了企业在社会网络大数据环境下,企业社会网络舆情给企业管理的各种挑战,并讨论了相应的应对管理策略。

思考题

1. 举例说明企业社会网络舆情给企业管理的挑战。
2. 列举企业在社会网络大数据环境下的舆情管理应对策略。

参 考 文 献

[1] Adams P H, Martell C H. Topic detection and extraction in chat. IEEE Int Conf Semantic Computing, 2008,581-588.

[2] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. Proc 14th ACM SIGKDD,2008,7-15.

[3] Andriole S J. Business impact of web2. 0 technologies, ACM Commun. 53 (12): 2010,67-79.

[4] Archak N,Ghose A,Ipeirotis P G. Deriving the pricing power of product features by mining consumer reviews. Management Science,2011,57(8): 1485-1509.

[5] Backstrom L,Huttenlocher D,Kleinberg J,Lan X Y. Group formation in large social networks: membership,growth,and evolution. Proc ACM SIGKDD,2006,44-54.

[6] Bambauer-Sachse S,Mangold S. Brand equity dilution through negative online word-of-mouth communication. Journal Retailing&Consumer Services, 2011, 18 (1): 38-45.

[7] Bizer C,Boncz P,Brodie M L, Erling O. The meaningful use of big data: four perspectives-four challenges. SIGMOD Record,2011,40(4): 56-60.

[8] Bouras C,Igglesis V,Kapoulas V, Tsiatsos T. A web based virtual community: functionality and architecture issues. Proc IADIS Int Conf,2004.

[9] Brian G,Kelly R. The importance of accurate road data for spatial applications in public health: customizing a road network. Int Journal Health Geographics,2009,8: 24.

[10] Cantador I,Konstas I,Jose J M. Categorising social tags to improve folksonomy-based recommendations,Web Semantics: Science. Services&Agents on World Wide Web,2011,9(1): 1-15.

[11] Carter,Craig R,Rogers, Dale S. A framework of sustainable supply chain management: moving toward new theory. Int Journal Physical Distribution & Logistics Management,2008,38(5): 360-387.

[12] Cebi S. A quality evaluation model for the design quality of online shopping websites. Electronic Commerce Research & Applications,2013,12(2): 124-135.

[13] Chen L，Tokuda N，Nagai A. A new differential LSI space-based probabilistic document classifier. Information Processing Letters，2003，88(5)：203-212.

[14] Chen Y，Fay S，Wang Q，The role of marketing in social media：how online consumer reviews evolve，Journal Interactive Marketing，2011，25(2)：85-94.

[15] Chen Y，Xie J. Online consumer review：word-of-mouth as a new element of marketing communication mix. Management Science，2008，54(3)：477-491.

[16] Chevalier J A，Mayzlin D. The effect of word of mouth on sales：online book reviews. Journal Marketing Research，2006，43(3)：345-354.

[17] Clark B H，Montgomery D B. Managerial identification of competitors. Journal Marketing，1999，67-83.

[18] Dellarocas C，Zhang X M，Awad N F. Exploring the value of online product reviews in forecasting sales：case of motion pictures. Journal Interactive Marketing，2007，21(4)：23-45.

[19] Dempsey J X，Flint L M. Commercial data and national security. George Washington Law Review，2004，27(6).

[20] Fearn-Banks K. Crisis communications：a casebook approach. Routledge，2010.

[21] Firan C，Nejdl W，Paiu R. The benefit of using tag based profiles. Proc 2007 Latin American Web Conf，2007，32-41.

[22] Ghosh R，Lerman K. Predicting influential users in online social networks. arXiv preprint arXiv：2010，1005-4882.

[23] Go A，Bhayani R，Huang L. Twitter sentiment classification using distant supervision. Tech Report，Stanford，2009，1-12.

[24] Goyal A，Bonchi F，Lakshmanan L V. Learning Influence Probabilities In Social Networks. Proc 3st ACM IntConf Web Search & Data Mining，2010，207- 217.

[25] Gu B，Ye Q. First step in social media—measuring the influence of online management responses on customer satisfaction. Production & Operations Management，2014，23(4)：570-582.

[26] He Y，Zhou D. Self-training from labeled features for sentiment analysis. Information Processing & Management，2011，47：606-616.

[27] Hersher R. Internet data miners' strike disease detection gold. Nature，2012，18(2)：185.

[28] Inouye D. Multiple post microblogsummarization. REU Research Tech Report，2010，1：34-40.

[29] Jones C，Hesterly W S，Borgatti S P. A general theory of network governance：exchange conditions and social mechanisms. Academy of Management Review，1997，22(4)：911-945.

[30] Keller K L. Brand synthesis: the multidimensionality of brand knowledge. Journal Consumer Research, 2003, 29(4): 595-600.

[31] Ketchen D J, Snow C, Hoover V L. Research on competitive dynamics: recent accomplishments and future challenges. Journal Management, 2004, 30 (6): 779-804.

[32] Khoo K B, Mitsuru I. Emerging topic tracking system, Advanced Issues of eCommerce and Web-Based Information Systems. 2001, 2-11.

[33] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines. Journal Machine Learning Research, 2005, 37-53.

[34] Ko H C. The determinants of continuous use of social networking sites. Electronic Commerce Research & Applications, 2013.

[35] Leitner P, Grechenig T. Collaborative shopping networks: sharing the wisdom of crowds in e-commerce environments. Proc 21st Bled eConf, Bled, Slovenia, 2008, 321-335.

[36] Li A, Shi Y, He J. MCLP-based methods for improving bad catching rate in credit cardholder behavior analysis. Applied Soft Computing, 2008, 3: 1259-1265.

[37] Li Q, Guo X, Zhao X, Bai X. Weekdays or weekends exploring the relationships between microblog posting patterns and addiction, ICIS, 2013.

[38] Li X, Snoek C M, Worring M, Koelma D, Smeulders A. Bootstrapping visual categorization with relevant negatives. IEEE Trans Multimedia, 2013, 15 (4): 933-945.

[39] Li X, Snoek C M, Worring M, Learning social image tag relevance by neighbor voting, IEEE Trans Multimedia, 2009, 11(7): 1310-1322.

[40] Li X, Snoek C M, Worring M, Smeulders A. Harvesting social images for bi-concept search. IEEE Trans Multimedia, 2012, 14(4): 1091-1104.

[41] Liang X. An effective method of pruning support vector machine classifiers. IEEE Trans Neural Networks, 2010, 21(1): 26-38.

[42] Liang X, Chen H, Yang J. A method of detecting and monitoring abnormal internet information, 美国专利(已授权), US 8185537, 2012.

[43] Liang X, Chen R, Guo X. Pruning support vector machines without altering performances. IEEE Trans Neural Networks, 2008, 19(10): 1792-1803.

[44] Liang X. Ni Z. Hyperellipsoidal statistical classifications in a reproducing kernel Hilbert space. IEEE Trans Neural Networks, 2011, 22(6): 968-975.

[45] Liang X. Social computing for public firms based on web information. Int Conf Computer Science & Engineering, Xian, 2012, 400-402.

[46] Lin D. An information—theoretic definition of similarity. ICML, 1998, 98:

296-304.
[47] Liu Y J, Li Q Q. Superedge prediction—what opinions will be mined based on an opinion supernetwork model. Decision Support Systems, 2014, 64: 118-129.
[48] Machado A. Drivers of shopping online: a literature review. Proc IADIS Int Conf eCommerce, 2005, 236-242.
[49] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A H. Big data: the nest frontier for innovation, competition, and productivity, McKinsey Global Institute, Annual Report, 6, 2011.
[50] Meng L, Huang R, Gu J. A review of semantic similarity measures in wordnet. Int Journal Hybrid Information Technology, 2013, 6(1): 1-12.
[51] Moreno C J, Naranjo C R, Sierra M L, et al. Building a Virtual E-Commerce Community. Proc IADIS Int Conf e-Commerce, 2004, 495-500.
[52] Neviarouskaya A, Prendinger H, Mitsuru I M. SentiFul - a lexicon for sentiment analysis. IEEE Trans Affective Computing, 2011, 2(1): 22-36.
[53] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 69(2): 026113, 2004.
[54] NitzanI, Libai B. Social effects on customer retention. Journal Marketing, 2011, 75 (6): 24-38.
[55] Nowak A, Szamrej J, Latane B. From private attitude to public opinion: a dynamic theory of social impact. Psychological Review, 1990, 97(3): 362.
[56] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. LREC, 2010.
[57] Parra-Santander D, Brusilovsky P. Improving collaborative filtering in social tagging systems for the recommendation of scientific articles. IEEE Int Conf Web Intelligence & Intelligent Agent Technology, 2010, 1: 136-142.
[58] Poyry E, Parvinen P, Malmivaara T. Can we get from liking to buying behavioral differences in hedonic and utilitarian facebook usage. Electronic Commerce Research & Applications, 2013.
[59] Ramage D, Dumais S T, Liebling D J. Characterizing Microblogs with Topic Models. ICWSM, 2010, 10: 1-1.
[60] Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proc ACL, 2005, 43-48.
[61] Rogers E M. Diffusion of Innovation. Free Press, 1995.
[62] Ruohomaa S, Kutvonen L. Trust management survey. Proc 3rd Int Conf Trust Management, Springer, 2005, 77-92.
[63] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity

of short text snippets,Proc 15th Int Conf World Wide Web,2006,377-386.

[64] Schutze H. Automatic word sense discrimination. Computational Linguistics,1998,24(1): 97-123.

[65] Sharifi B P. Automatic microblogclassification and summarization. Tech Report, University of Colorado,2010.

[66] Sherchan W,Nepal S,Paris C. A survey of trust in social networks, ACM Computing Surveys,2013,45(4): 47.

[67] Smeulders A W M,Worring M, Santini S, et al. Content-based image retrieval at the end of the early years,Pattern Analysis and Machine Intelligence. IEEE Trans, 2000,22(12): 1349-1380.

[68] Stutzman F. An evaluation of identity-sharing behavior in social network communities. Int Digital&Media Arts,2006,3(1): 10-18.

[69] Talluri S,Baker R C, Sarkis J. A framework for designing efficient value chain networks. Int Journal Production Economics,1999,62(1): 133-144.

[70] Tian R Y,Liu Y J. Isolation, insertion, and reconstruction - three strategies to intervene in rumor spread based on supernetwork model. Decision Support Systems,2014,67: 121-130.

[71] Tsai W,Ghoshal S. Social capital and value creation: the role of intrafirm networks,Academy of Management Journal,1998,41(4): 464-476.

[72] Villanueva J,Yoo S, Hanssens D M. The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. Journal Marketing Research,2008,45(1): 48-59.

[73] Wang X,Zhang L,Li Xirong,Ma W. Annotating images by mining image search results. IEEE Trans Pattern Analysis & Machine Intelligence,2008,30(11): 1919-1932.

[74] Xiang R,Neville J,Rogati M. Modeling relationship strength in online social networks,Proc 19th Int Conf World Wide Web. ACM Press,2010,981-990.

[75] Xu K,Wang W,Ren J. Classifying consumer comparison opinions to uncover product strengths and weaknesses. Information Technologies,1,2012.

[76] Ye Q,Fang B. He W J. Can social capital be transferred cross the boundary of the real and virtual worlds - an empirical investigation of Twitter. Journal Electronic Commerce Research,2012.

[77] Ye Q,Law R,Gu B. The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human Behavior,2011,27(2): 634-639.

[78] Ye Q,Shi W,Li Y J. Sentiment classification for movie reviews in Chinese by

improved semantic oriented approach,System Sciences,HICSS,Proc39th Annual Hawaii IntConf,2006,3:53b-53b.
[79] Ye Q,Zhang Z,Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications,2009,36(3):6527-6535.
[80] Zarghami A,Fazeli S, Dokoohaki N, Matskin M. Social trust-aware recommendation system: A t-index approach,Proc IEEE Int Joint Conf Web Intelligence&Intelligent Agent Technology,Washington,2009,3:85-90.
[81] Zhang W,Xu H,Wan W. Weakness finder:find product weakness from Chinese reviews by using aspects based sentiment analysis,Expert Systems with Applications,2012,39(11):10283-10291.
[82] Zhang Z,Ye Q,Zhang Z. Sentiment classification of internet restaurant reviews written in Cantonese,Expert Systems with Applications,2001,38(6):7674-7682.
[83] Zhou G,Xu K Q,Liao S Y. Do starting and ending effects in fixed-price group-buying differ. Electronic Commerce Research&Applications,2013.
[84] Zhu F,Zhang X. Impact of online consumer reviews on sales:the moderating role of product and consumer characteristics. Journal Marketing,2010,74(2):133-148.
[85] 蔡登,卢增祥,李衍达. 信息协同过滤. 计算机科学,2002,29(6):1-4.
[86] 曹磊,陈薇娜,缪其浩,陈超. 大数据:数字世界的智慧基因. 文汇报,2011.
[87] 曹润. 基于用户联系强度及关注网络的中文微博社区发现研究. 硕士学位论文(导师:梁循),2013.
[88] 曹润,梁循,李亚平,李倩,朱浩然. 一种多模式微博信息动态排序方法,CN201210199559,2,2012.
[89] 陈华,梁循,阮进,互联网舆情关联分析系统的设计实现,信息检索与内容安全学术会议论文集,苏州,2007,45-49.
[90] 戴媛,姚飞. 基于网络舆情安全的信息挖掘及评估指标体系研究. 情报理论与实践,2009,6:873-876.
[91] 丁菊玲,勒中坚. 基于观点树的网络舆情危机预警方法. 计算机应用研究,2011,28(9):3501-3504.
[92] 杜智涛,谢新洲. 利用灰色预测与模式识别方法构建网络舆情预测与预警模型. 图书情报工作,2013,57(15):27-33.
[93] 桂斌,杨小平. 微博意群的情感倾向分析. 中文信息学报,2015 (已录用).
[94] 郭小钗,陈蓓蕾. 在线口碑效应的影响因素实证研究. 北京理工大学学报,2009,11(2):31-35.
[95] 郭晓科. 大数据. 北京:清华大学出版社,2013.
[96] 郝媛媛,叶强,李一军. 基于影评数据的在线评论有用性影响因素研究. 管理科学

学报，2010，78-88，96.
[97] 何佳，周长胜，石显锋. 网络舆情监控系统的实现方法. 郑州大学学报：理学版，2010，42(1)：82-85.
[98] 何静，郭进利，徐雪娟. 微博用户行为统计特性及其动力学分析. 情报分析与研究，2013，7：94-100.
[99] 胡百精. 公共传播经典译丛. 北京：中国传媒大学出版社，2014.
[100] 胡百精. 互联网与现代传播. 中国传媒大学学报，2014a，(2)：40-46.
[101] 胡百精. 新媒体与社会信任. 国际公关，2013，5：44.
[102] 胡昌平. 信息管理科学导论. 科学技术文献出版社，1995.
[103] 胡大立. 基于价值网模型的企业竞争战略研究. 中国工业经济，2006，222(9)：87-93.
[104] 惠志斌. 面向网络传播形态的公共危机信息预警模式研究. 图书情报工作，2012，56(2)：71-75.
[105] 吉祥. 基于观点挖掘的网络舆情信息分析. 现代情报，2010，30(11)：46-49.
[106] 金淳，张一平. 基于 Agent 的顾客行为及个性化推荐仿真模型. 系统工程理论与实践，2013，33(2)：463-472.
[107] 康伟. 基于 SNA 的突发事件网络舆情关键节点识别——以 7・23 动车事故为例. 公共管理学报，2012，9(3)：101-111.
[108] 雷雳. 青少年"网络成瘾"干预的实证基础. 心理科学进展，2010，6：791-797.
[109] 雷雳. 青少年网络成瘾探析. 心理发展教育，2010，5：554-560.
[110] 李弼程，王瑾，林琛. 基于直觉模糊推理的网络舆情预警方法. 计算机应用研究，2010，9：3312-3315.
[111] 李纲，陈璟浩. 突发公共事件网络舆情研究综述. 图书情报知识，2014，2：111-119.
[112] 李季梅，陈宁，陈安，武艳南. 突发事件的网络舆情监测与恐慌度量系统. 中国科技资源导刊，2009，41(2)：62-67.
[113] 李建平，徐伟宣，刘京礼，石勇. 消费者信用评估中支持向量机方法研究. 系统工程，2004，22(10)：35-39.
[114] 李实，陆光. 修正中文评论挖掘中产品特征词序的实验研究. 科学技术与工程，2012，12(21).
[115] 李实，叶强，李一军，Law R. 挖掘中文网络客户评论中的产品特征方法研究. 管理科学学报，2009，12(2)：142-152.
[116] 梁循，马跃峰，杨小平，林航. 基于移动终端及 ZigBee 组件的应急信息系统. 电子科技大学学报，2013，15(6)：16-19，29.
[117] 梁循，申华，曹润. 一种面向微博的全新突发事件发现方法. CN201210250175，9，2012.

[118] 梁循.数据挖掘算法与应用.北京:北京大学出版社,2006.

[119] 梁循,杨小平,申华.社会化商务理论与实践.北京:清华大学出版社,2014.

[120] 梁循,杨小平,周小平.面向微博大数据的社会计算及其应用.北京:清华大学出版社,2014.

[121] 林光.集团公司生命周期系统的管理.北京:清华大学出版社,2005.

[122] 林敏.网络舆情:影响因素及其作用机制研究.2013.

[123] 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展,2009,19(1):1-15.

[124] 刘京礼,李建平,徐伟宣,石勇.信用评估中的鲁棒赋权自适应 Lp 最小二乘支持向量机方法.中国管理科学,2010,18(5):28-33.

[125] 刘晓亮.电子商务信誉评价乱象与解决途径.电子商务,2013,(2):84-85.

[126] 刘毅.网络舆情研究概论.天津:天津人民出版社,2007.

[127] 刘毅.略论网络舆情的概念,特点,表达与传播.理论界,2007,1:11-12.

[128] 娄策群,周承聪.信息生态链:概念,本质和类型.图书情报工作,2007,51(9):29-32.

[129] 路荣,项亮,刘明荣.基于隐主题分析和文本聚类的微博客新闻话题发现研究.第六届全国信息检索学术会议,2010,291-298.

[130] 吕嘉.重新理解社会存在决定社会意识.哲学动态,(6):2001:6-9.

[131] 毛基业.管理信息系统.北京:清华大学出版社,2011.

[132] 邱云飞,王琳颖,邵良杉,郭红梅.基于微博短文本的用户兴趣建模方法.计算机工程,40(2):2014:275-279.

[133] 申华,梁循.基于 SIFT 与 SVM 图像异常检测的应急管理模型.哈尔滨工业大学学报,2015(已录用).

[134] 施国良,程楠楠.Web 环境下产品评论挖掘在企业竞争情报中的应用.情报杂志,30(11):2011.

[135] [美]施拉姆.波特传播学概论.陈亮,周立方,李启,译.北京:新华出版社,1984.

[136] 施晓菁,梁循,曹润,周晨曦.基于兴趣分析的微博博主社区分类方法.CN 201210250181.4,2012.

[137] 施晓菁,梁循,孙晓蕾.基于在线评级和评论的评价者效用机制研究.中国管理科学,2015 (已录用).

[138] 史波.公共危机事件网络舆情内在演变机理研究.情报杂志,2010,29(4):41-45.

[139] 史学敏,张闯.基于微博用户行为的时区预测.中国科技论文在线,2011.

[140] 孙帅.突发事件网络舆情管理机制研究.苏州大学硕士学位论文.2014.

[141] 涂子沛.大数据革命.桂林:广西师范大学出版社,2012.

[142] 万飞,梁循,赵溪,曹润,梁霞,付虹蛟,周晨曦.一种基于文本内容的垃圾微博过滤方法.CN 201210199582.8,2012.

[143] 王超,李楠,李欣丽,梁循.文本情感倾向性分析用于金融市场波动率的研究.中文信息学报,23(1):2009:95-99.

[144] 王国华,张剑,毕帅辉.突发事件网络舆情演变中意见领袖研究——以药家鑫事件为例.情报杂志,30(12):2012:1-5.

[145] 王来华.舆情研究概论.天津:天津社会科学院出版社,2003.

[146] 王实,高文,李锦涛.基于隐马尔可夫模型的兴趣迁移模式发现.计算机学报,24(2):2012:152-157.

[147] 王伟.基于企业基因重组理论的价值网络构建研究.中国工业经济,203(2):2005:58-65..

[148] 王伟.王洪伟,孟园.协同过滤推荐算法研究:考虑在线评论情感倾向.系统工程理论与实践,34(12):2014:3238-3249.

[149] 维克托·迈尔-舍恩伯格,肯尼思·库克耶.大数据时代.杭州:浙江人民出版社,2013.

[150] 肖渡,沈群红.合作网络形成的理论探讨及其意义.管理工程学报,14(4):2000:69-73.

[151] 肖勇.现代信息科学结构与内涵分析.图书情报工作,2001,12:9-14.

[152] 肖玉芝,赵海兴.基于超图理论的在线社会网络用户行为分析.计算机应用与软件,2014,31(7):50-54.

[153] 谢科范,赵湜,陈刚,蔡文静.网络舆情突发事件的生命周期原理及集群决策研究.武汉理工大学学报:社会科学版,2010,23(4):482-486.

[154] 谢丽星,周明,孙茂松.基于层次结构的多策略中文微博情感分析和特征抽取.中文信息学报,2012,26(1):73-83.

[155] 杨瑞龙,冯健.企业间网络及其效率的经济学分析.江苏社会科学,2004,3:53-58.

[156] 杨源,马云龙,林鸿飞.评论挖掘中产品属性归类问题研究.中文信息学报,2012,26(3).

[157] 叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究.信息系统学报,2007,79-91.

[158] 曾春,邢春晓,周立柱.个性化服务技术综述.软件学报,2002,13(10):1952-1996.

[159] 曾润喜.网络舆情突发事件预警指标体系构建.情报理论与实践,2010,1:77-80.

[160] 曾润喜,徐晓林.网络舆情对群体性突发事件的影响与作用.情报杂志,2010,29(12):1-4.

[161] 曾润喜,徐晓林.网络舆情突发事件预警系统、指标与机制.情报杂志,2009,(11):52-54.

[162] 詹志建,杨小平.基于改进遗传算法的中文短文本相似度计算.中文信息学报,

2015 (已录用).
[163] 詹志建.中文短文本相似度计算关键技术研究.博士学位论文(导师：杨小平),2014.
[164] 张海燕,梁循,周小平.针对有向图的局部延展的重叠社区发现算法.数据采集与处理,2014.
[165] 张海燕,孟祥武.基于社会标签的推荐系统研究.情报理论与实践,2012,5：103-107.
[166] 张克生.国家决策：机制与舆情.天津：天津社会科学院出版社,2004.
[167] 张亮.Web数据挖掘在群体性事件预警系统中的应用.光盘技术,2009,6：26-28.
[168] 张秀伟,何克清,王健,刘建晓.Web服务个性化推荐研究综述.计算机工程与科学,2013,35(9)：132-140.
[169] 张一文,齐佳音,方滨兴,李欲晓.基于贝叶斯网络建模的非常规危机事件网络舆情预警研究.图书情报工作,2012,56(2)：76-81.
[170] 张振国,宋薇,李婧.基于序列模式挖掘的社交网络用户行为分析.现代情报,2013,33(3)：56-60.
[171] 赵军力,梁循.基于TrTS取样的股票收益率RV测度的改进.中国管理科学,2015 (已录用).
[172] 郑小平.在线评论对网络消费者购买决策影响的实证研究.中国人民大学硕士学位论文(导师：毛基业),2008.
[173] 周伟恒.基于生命周期理论的企业舆情管理研究,华中师范大学硕士学位论文,2013.
[174] 周小平,梁循,张海燕.基于R-C模型的微博用户社区发现,软件学报.2014,25(12)：2808-2823.
[175] 周晓飞,石勇.基于数据挖掘的金融信用评估概述.中国管理学年会论文集,2009.
[176] 周珍妮,陈碧荣.Web 3.0：全新的互联网时代.图书情报论坛,2009.
[177] 朱浩然,梁循,马跃峰,纪阳,李启东.金融领域中文微博情感分析.上海：中国管理学年会论文集,2013.
[178] 祝效国,叶强,李一军.企业技术创新的采纳、扩散与内化机制研究综述.信息系统学报,2009,3(2)：66-76.
[179] 邹本友.基于用户信任和张量分解的社会网络推荐.软件学报,2014,25(12)：2852-2864.